Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia

Xiaowei Zhu, Bo Zhou, Reenal Pattni, Kelly Gleason, Chunfeng Tan, Agnieszka Kalinowski, Steven Sloan, Anna-Sophie Fiston-Lavier, Jessica Mariani, Dmitri Petrov, Ben A. Barres, Laramie Duncan, Alexej Abyzov, Hannes Vogel, Brain Somatic Mosaicism Network, John V. Moran, Flora M. Vaccarino, Carol A. Tamminga, Douglas F. Levinson and Alexander E. Urban

Retrottransposons can cause somatic genome variation in the human nervous system, which is hypothesized to have relevance to brain development and neuropsychiatric disease. However, the detection of individual somatic mobile element insertions presents a difficult signal-to-noise problem. Using a machine-learning method (RetroSom) and deep whole-genome sequencing, we analyzed L1 and Alu retrotransposition in sorted neurons and glia from human brains. We characterized two brain-specific L1 insertions in neurons and glia from a donor with schizophrenia. There was anatomical distribution of the L1 insertions in neurons and glia across both hemispheres, indicating retrotransposition occurred during early embryogenesis. Both insertions were within the introns of genes (CNNM2 and FRMD4A) inside genomic loci associated with neuropsychiatric disorders. Proof-of-principle experiments revealed these L1 insertions significantly reduced gene expression. These results demonstrate that RetroSom has broad applications for studies of brain development and may provide insight into the possible pathological effects of somatic retrottransposition.

About 45% of the human genome is composed of mobile elements (MEs), which include cut-and-paste DNA transposons and 'copy-and-paste' retrottransposons (acting via RNA intermediates). Most of these elements are inactive, but three classes of active retrottransposons—human-specific L1 (L1Hs), AluY and SVA (SINE/VNTR/ALU)—can undergo retrotransposition via target-primed reverse transcription (TPRT). De novo retrottransposition events in both germline and somatic tissue can create mobile element insertion (MEI) mutations and precipitate genomic structural rearrangements. L1 (31 cases) and Alu (over 70 cases) germline mutations have been reported for monogenic diseases. Specific somatic MEIs have been detected at high levels of mosaicism in some human cancers (sometimes in more than 25% of tumor cells) and at lower levels in human brain (for example, ~1% of cells for each examined brain region). Dysregulation of retrottransposition has been hypothesized to contribute to neurogenetic diseases and elevated L1 activity is proposed to be associated with neuropsychiatric disorders. Somatic L1 retrottransposition events also have been reported to occur in neural precursor cells during early human and mouse embryogenesis, and their regional distributions have been used to trace neuronal cell lineages.

Because individual somatic MEIs are present in a small proportion of brain cells, standard whole-genome sequencing (WGS) is facing a difficult signal-to-noise problem. Studies reporting on brain somatic MEIs have addressed this problem using either a capture approach, such as retrottransposon capture sequencing from bulk brain tissue, or single-cell-based approaches (because a somatic MEI is heterozygous within each mutated cell), which include single-cell retrottransposon capture sequencing, single-cell L1 insertion profiling, single-cell WGS (sc-WGS) and single-cell L1-associated variant sequencing. A drawback of these methods is the occurrence of sequencing artifacts via chimeric DNA molecules that arise from the high numbers of PCR cycles (capture) or from the massive enzymatic whole-genome amplification. Furthermore, it is very expensive to apply sc-WGS to hundreds of cells derived from multiple regions of an individual brain sample. Lastly, MEI detection using all WGS approaches relies on unique mapping highly repetitive sequencing reads to the genome, which remains a challenging task.

Here, we developed a new analytic method, RetroSom, to detect somatic L1 and Alu MEIs in deep (200× coverage) WGS data from sorted fractions of brain cells. Using RetroSom, we discovered and validated two individual somatic L1 insertions in the human brain, which were absent from control tissues and present in similar cellular proportions and anatomical distributions in glia and neurons in both brain hemispheres. This approach is not susceptible to amplification artifacts and is more cost-effective than current sc-WGS technologies for MEI detection.
For WGS, we used genomic DNA extracted from sorted cells (typically more than 100,000 cells per cell-type fraction) from one anatomical location for each brain (Fig. 1a,b). MEI detection was then based on two types of sequencing reads (Fig. 1c): (1) split reads (SR), which capture the MEI insertion point such that part of the read maps to the ME consensus sequence and the other part maps to the unique flanking reference sequence at the new genomic location; and (2) paired-end (PE) reads, where one read maps to the ME consensus and the other to the unique flanking sequence. In both cases, the unique sequence localizes the MEI in the genome. Existing algorithms based on these principles can detect germline MEIs\(^{16}\), somatic MEIs in single cells\(^{11,13}\) and MEIs carried by a high subclonal fraction of tumor cells (\(>25\%\))\(^{14}\), but they require many supporting reads (for example, \(\geq 5\)) per MEI for reliable detection. Lowering the detection threshold (for example, to \(\leq 2\) supporting reads) leads to overwhelming numbers of false positives, likely due to experimental noise and alignment errors\(^{15}\). For example, using one supporting read in WGS data at 50x genomic coverage, we should detect \(\geq 50\%\) of MEIs that are present in \(\geq 96\%\) of cells. However, using a standard MEI algorithm, RetroSeq\(^{14}\), to detect calls with one supporting read, yielded \(~59,900\) (95% confidence interval (CI): 55,100–64,700) false-positive MEI detections (Fig. 1d and Extended Data Fig. 1a).

RetroSom integrates RetroSeq (for mapping of reads to ME or reference sequence) with a learning transfer model trained on evolutionarily recent germline MEIs to detect low-level somatic MEIs. We separately analyzed neurons (NeuN\(^+\)) and nonneuronal (NeuN\(^-\), mostly glial) cells derived from five adult human postmortem brains; one elderly adult (‘AIS’), two schizophrenia–control pairs (Dallas Brain Collection) and neurons (CD45\(^-\)/HepaCAM\(^-\)/Thy1\(^+\)) and astrocytes (CD45\(^-\)/Thy1\(^-\)/O4\(^-\)/HepaCAM\(^-\)) from one fetal brain (‘F1’; Supplementary Fig. 1 and Supplementary Table 1). We collected tissue from the superior temporal gyrus (STG) of adult brains because of ample availability of tissue and relevance to schizophrenia in neuroimaging studies\(^{16}\), cortical tissues from fetal brain, and matched heart or fibroblast control tissue. We sequenced extracted genomic DNA from each specimen to 200x whole-genome coverage (Fig. 1a,b). Additional data used for algorithm development are described in Supplementary Table 2.

**Results**

Optimization of somatic MEI detection with machine learning.

We trained RetroSom using polymorphic germline MEIs selected from Illumina Platinum Genomes WGS data\(^{20}\) for 17 members of a three-generation pedigree (Fig. 1e and Supplementary Table 2). We assumed that recent germline MEIs would produce high-confidence non-reference calls that segregate in a Mendelian fashion. We excluded genomic regions of poor mapping quality based on preestablished criteria, including telomeric or centromeric repeats, segmental duplications, gaps or reference MEI insertions of the same type and on the same strand, totaling 21% of the genome for detection of Alu or 24% for L1. We also removed regions with abnormal sequencing depth and supporting reads with low sequence complexity. We defined true-positive MEIs based on their inheritance pattern. Criteria for false MEI calls (likely artifacts) were fewer than three supporting reads in offspring and missing in both parents. We detected non-reference true-positive insertions including, on average, 89 L1 and 467 Alu for each offspring (Extended Data Fig. 1c). We then chose 16–28 sequence features for each of the four supporting-read classes (L1 and Alu elements, PE and SR for each element) to help distinguish true retrotransposition of evolutionary young and active retrotransposons from noise generated by old and inactive elements (Supplementary Table 3). We excluded several features to help generalization from germline to somatic MEIs including: (1) the number of supporting reads (used as a selection criterion for true-positive MEIs); (2) features specific to individual elements (for example, unique single-nucleotide polymorphisms (SNPs)/indels, unlikely to be shared by other families); (3) features specific to sequencing conditions (for example, sequencing read length); and (4) chromosomal location (for example, positional bias in germline MEIs could be due to natural selection or genetic drift and irrelevant to somatic MEIs\(^{11}\)).

We developed a machine-learning algorithm using the above features to classify true or false L1 or Alu supporting reads (Extended Data Fig. 1d,e). We tested logistic regression (with and without regularization), random forest\(^{21}\) and naïve Bayes classifiers, using 11x cross-validation (training on ten offspring, testing on the eleventh). In imbalanced training data, where the negatives outnumber the positives, a relatively high level of false positives could still yield excellent specificity (true negatives/(true negatives + false positives)) but poor precision (true positives/(true positives + false positives)). Thus, we used precision as a better index in the context of our project. The random forest model, an ensemble method that combines multiple decision trees from data subsampling, performed best, with an area under the precision-recall curve of 0.965 (95% CI: 0.959–0.971; Extended Data Fig. 1f,g). The most important differentiating features were sequence homology to the L1Hs or AluY consensus (Fig. 1f), L1Hs-specific SNPs (Fig. 1g)\(^{22}\) and exclusion of Alu calls with flanking sequences from the putative source locations (‘transduction’, which can occur with L1, but not Alu, retrotransposition events; Fig. 1h)\(^{24}\).

**Performance evaluation in independent test datasets.** We tested RetroSom in several independent WGS datasets. Data from clonally expanded fetal brain cells\(^{25}\) confirmed that more than two supporting reads were necessary for high precision (L1: 99.97%; Alu: 99.99%) with adequate sensitivity (L1: 49.5%; Alu: 82.52%; Fig. 2a, Extended Data Fig. 2a and Supplementary Note 1). We also identified one somatic L1 insertion with features suggesting an insertion arising by an internal priming event\(^{26}\), a rare
endonuclease-independent retrotransposition process\textsuperscript{27} or an unknown alternative mechanism (Extended Data Fig. 3 and Supplementary Note 2). In addition, Illumina sequencing libraries prepared using a PCR-based method (approximately ten cycles) yielded 30–1,000% more false MEIs than PCR-free libraries, many due to sequencing errors around low-complexity regions from PCR.

---

**Articles**

**Nature Neuroscience**

VOL 24 | FEBRUARY 2021 | 186–196 | www.nature.com/natureneuroscience
polymerase slippage (Supplementary Fig. 2). However, RetroSom removed all false MEIs, yielding similar sensitivities for the two library types (L1: ~70%; Alu: ~86%; Fig. 2b, Extended Data Fig. 2b and Supplementary Note 3). We note that these sensitivity measurements may be an overestimate also because L1 (and presumably Alu) ‘transposon-in-transposon’ insertions are challenging to detect, in principle, with standard short-read sequencing6.

We further benchmarked RetroSom using a genome-mixing experiment. We pooled DNA from six human genomes (for which we called high-confidence germline MEIs from available Illumina sequencing data) in precise proportions of 0.2–25% with HapMap sample NA12878 (whose germline MEIs are generally established). We sequenced the pool (and NA12878 separately as a control) to 200× coverage and called MEIs using RetroSom. A heterozygous germline MEI present in only one of the six genomes will appear as a mosaic MEI in the WGS data from the DNA mix, with few (if any) supporting reads. RetroSom L1 detection sensitivities were 0 at mixing proportions of 0.04% and 0.2%, 0.16 at 1%, 0.67 at 5% and 0.90 at 25%, with no false positives (Fig. 2c,d). Detection rates were higher for RetroSeq alone (0.32 for 1%) or using RetroSom and relying on just one supporting read (0.48 for 1%), but also yielded 4,316 and 584 false positives, respectively (Fig. 2e). Sequencing depth, when computationally varied from 50× to 400×, linearly predicted detection sensitivity (especially for MEIs mixed in low proportions), but not precision (Fig. 2c–e). RetroSom was more sensitive and less precise for Alu, detecting five Alu insertions at 0.2% mosaicism with five false positives (Extended Data Fig. 2c–e). This excess of false positives could be due to the higher abundance of genomic Alu sequences with <5% sequence divergence from the active consensus sequence (26,720 Alu sequences versus 1,531 L1s). Thus, using 200× WGS data, these mixing controls indicate that RetroSom can detect most L1 and Alu MEIs at >5% mosaicism, one-sixth with 1% mosaicism and <1/100 with <0.2% mosaicism.

**Discovery and validation of somatic mobile element insertions.**

We applied RetroSom to 200× WGS data from sorted neurons, sorted glia and a control tissue from A1S, F1 and the two Dallas schizophrenia–control pairs; we then called somatic MEIs (≥2 high-confidence supporting reads in either brain fraction but none in the corresponding control). As above, we again excluded 21% of the genomic sequence from analysis for Alu and 24% for L1 MEIs. There were 0–3 putative somatic L1 and 0–13 putative somatic Alu calls per fraction (Supplementary Table 4). We selected MEIs for validation by blinded manual inspection with a new visualization

---

**Fig. 2 | Benchmarking in independent test datasets.**

**a.** Performance in the detection of germline L1 insertions from clonally expanded fetal brain cell sequencing data. Gray, clones from donor 316 sequenced with WGA (316 WGA; n = 10 clones); brown, the rest of the 316 datasets (316 noWGA; n = 5 clones); blue, clones from donor 320 (n = 53 clones). The boundaries of the box plots indicate the 25th (above) and 75th (below) percentiles, and the black line within the box marks the median. Whiskers above and below the box indicate the 10th and 90th percentiles.

**b.** Performance in the detection of germline L1 insertions from sequencing libraries prepared with or without PCR. Light blue and green, PCR-free libraries for sample ‘heart’ (light blue circle; n = 1 library) and ‘neuron’ (light green triangle; n = 1 library); dark blue and green, PCR-based libraries for ‘heart’ (dark blue circle, n = 6 libraries) and ‘neuron’ (dark green triangle; n = 6 libraries).

**c–e.** Performance in the detection of somatic MEIs simulated by six genomic DNA samples at proportions of 0.04% to 25% with that of NA12878, at various sequencing depths (gray, 50×; brown, 100×; blue, 200×; green, 400×). Similar performances were observed in the detection of Alu insertions (Extended Data Fig. 2).
tool (RetroVis), following a checklist of screening criteria (Extended Data Fig. 4). We excluded most L1 and all Alu putative insertions, which generally resulted from misalignment of the reads mapped to the flanking sequence, germline insertions and potential PCR duplicates or chimeras (Supplementary Table 4). Two brain L1 insertions (L1-1 and L1-2), both from the same schizophrenia donor brain (ID 12004), fulfilled all criteria and were subjected to in-depth investigation (Extended Data Fig. 5 and Supplementary Table 1). Additional germline variants detected in the donor samples are described in Supplementary Note 4.

We validated both L1 insertions following guidelines established by the Brain Somatic Mosaicism Consortium and the MEI research community. We quantified mosaicism levels using droplet digital PCR (ddPCR), determined the genomic DNA/L1 junction sequences by nested PCR and characterized the full-length sequences (single-base resolution) by overlap extension PCR, using genomic DNA from the site of discovery (right STG) as the input (Extended Data Figs. 5–7 and Supplementary Note 5). L1-1 was discovered with two high-quality PE supporting reads in neurons, covering the upstream and downstream junctions (Fig. 3a and Supplementary Fig. 3a). Estimated mosaicism levels were 0.72% of neurons (95% CI: 0.50–0.94%), 0.54% of glia (95% CI: 0.40–0.67%) in the discovery region and 0% in fibroblasts (eight technical replicates; Fig. 3b and Extended Data Fig. 6b). The full insertion sequence demonstrated four hallmarks of in vivo L1 retrotransposition (Fig. 3c and Extended Data Fig. 6c): (1) the endonuclease cleavage site is 5′-TTTT/AA-3′, similar to the degenerate consensus motif 5′-TTTT/AA-3′ (ref. 32); (2) consistent with the common 5′ truncation of new L1 insertions, L1-1 is a 384-bp 3′ fragment of the L1 consensus, with a poly(A) tail of ~35 bp that is in the 18th percentile when compared to the lengths of tails of the 22 de novo disease-causing L1 retrotranspositions with known poly(A) lengths (Extended Data Fig. 8c,d) and exhibits a short region of microhomology at the 5′ genomic DNA/L1 sequence junction; (3) we confirmed a 15-bp target site duplication (TSD), as expected with TPRT retrotransposition; (4) L1-1 carries the diagnostic ACA allele at base 5927–5929, the G allele at base 6012 and no other mismatches to the L1Hs consensus sequence, indicating that the source element is from the youngest L1hs subfamily, L1-Hs-Ta (Extended Data Fig. 6c)

L1-2 was discovered with three supporting reads, including an SR spanning the upstream junction (Fig. 3d and Supplementary Fig. 3b). Estimated mosaicism levels were 1.2% of neurons (95% CI: 1.0–1.4%), 0.53% of glia (95% CI: 0.46–0.60%) and 0% in fibroblasts (eight technical replicates; Fig. 3e and Extended Data Fig. 7b). The endonuclease site is 5′-CTTT/AA-3′, and the sequence contains a 418-bp 3′ fragment of the consensus sequence, a poly(A) tail of ~25 bp (ranked in the 14th percentile; Extended Data Fig. 8c,d), a 4-bp 5′ microhomology and a 6-bp TSD (Fig. 3f). L1-2 also belongs to the L1-Ta subfamily, with one mismatch when compared to the L1Hs consensus sequence (Extended Data Fig. 7c).

Spatial occurrence of somatic L1 retrotransposition in neurons and glia. Previous studies detected individual L1 insertions in neurons, with narrow or broad distributions in one hemisphere of the brain. Here, we detected L1-1 and L1-2 in neurons and glia from 24 brain regions, from symmetrical sites across both hemispheres (Fig. 4 and Extended Data Fig. 8a). L1-1 was detected in neurons from all 24 regions (0.05–2.46% mosaicism), and glia from 17 regions (0.05–14.4%: Fig. 4a,c), including the putamen in the basal ganglia and the cerebellum, with the maximum mosaicism level detected in the left STG (neurons: 1.1% (95% CI: 0.2–2.4%); glia: 14.4% (95% CI: 13.0–15.9%)). L1-2 was absent in specimens from the prefrontal cortex, putamen and cerebellum. It was detected in 12 of 24 regions, all in the cerebral cortex (neurons: 0.1–1.4%; glia: 0.07–1.1%; Fig. 4b,d), with the maximum mosaicism level detected in the right occipital cortex distal to the STG. For both insertions, mosaicism levels were similar in neurons and glia from the same regions (Spearman ρ = 0.77, P = 1.3 × 10−10; Extended Data Fig. 8b). We further developed a droplet-based full-length PCR approach to verify the full-length post-integration allele for L1-1 from glia in the left occipital cortex proximal to STG (L1-1, mosaicism: 3.8%) and left STG (L1-2, mosaicism: 14.4%), and for L1-2 from neurons in the right occipital cortex distal to STG (R0D; mosaicism: 1.3%; Supplementary Note 5).

Dysregulation of gene expression by L1 insertion. L1-1 is inserted into an intron of CNNM2 (antisense strand), while L1-2 is in an intron of FRMD4A (sense strand). More precisely, L1-1 is inserted within a 2.6-kb putative transcriptional regulatory element ENSRNO0000032826 (Ensembl v89; Fig. 5), as determined by transcription factor binding and epigenetic marker patterns. L1-1 is also inserted in a broad linkage disequilibrium (LD) region surrounding AS3MT and CNNM2, where genome-wide significant evidence for association was reported for schizophrenia and several other traits (Fig. 5, Extended Data Fig. 9 and Supplementary Table 5).

CNNM2 and FRMD4A are expressed in many tissues, with higher levels in the brain (Supplementary Note 6). Tissue culture studies show that intronic L1 insertions, either on the sense or antisense strand relative to the transcriptional orientation of the gene, can alter or disrupt gene expression (for example, by inhibiting transcription elongation, altering splicing, terminating transcription prematurely or modifying local chromatin structure)34. The strength of the effect depends on insertion position within the intron, insertion length, strand and splicing or polyadenylation sites within the insertion.

Using an enhanced green fluorescent protein (EGFP) reporter ‘Gint’ in cell culture, we conducted proof-of-principle experiments to gauge the potential effects of L1-1 and L1-2 on gene expression by cloning the full-length insertions (with flanking sequences) into a constitutively spliced intron in the antisense or sense strand, respectively, of the EGFP locus (Fig. 6a and Extended Data Fig. 5b). Control reporters were generated for the two flanking sequences lacking an L1 insertion. In blinded experiments, we co-transfected each of the modified L1 inserts.

Fig. 3 | Discovery and experimental validation of somatic L1-1 and L1-2. a, L1-1 was identified by RetroSom with two supporting sequencing reads, and the insertion is in the antisense strand of an intron of CNNM2. Blue, read that maps to the flanking sequence; red, mate read that maps to the L1 consensus; pT, poly(T) tail of L1-1. b, ddPCR targeting the L1-1 upstream flanking junction confirms the insertion is present in both neurons (0.72%) and glia (0.54%), but absent in the fibroblast and NA12878. c, With Sanger sequencing of the 5′ and 3′ junctions, we confirmed the L1 insertion had an endonuclease cleavage site 5′-TTTT/CA-3′ and a 15-bp TSD. The inserted L1 element was CTTA (ranked in the 14th percentile; Extended Data Fig. 8c,d), a 4-bp 5′ microhomology and a 6-bp TSD (Fig. 3f). L1-2 also belongs to the L1-Ta subfamily, with one mismatch when compared to the L1Hs consensus sequence (Extended Data Fig. 7c).
GFPs expressing Gint reporters with a red fluorescent protein (RFP) expressing control plasmid ‘Rint’ into HeLa cells and measured the level of fluorescence (Fig. 6b,d,e). Compared to controls, L1-1 (anti-sense) reduced green fluorescence by 28% (95% CI: 20–35%, Welch’s two-sided test, $t = -6.2$, $df=1,210.1$, adjusted $p=8\times10^{-9}$), whereas L1-2 (sense) reduced green fluorescence by 39% (95% CI: 33–45%, $t = -6.2$, $df=1,210.1$, adjusted $p=8\times10^{-9}$).
Including the intronic length as a covariate, the difference in fluorescence remained significantly correlated for insertion versus control assay ($t = -9.27, df = 2,321, \text{adjusted } P = 4 \times 10^{-19}$). The strength of the effect by L1-2 was also significantly higher than by L1-1 ($t = 4.12, df = 1,027.7, \text{adjusted } P = 3 \times 10^{-4}$), possibly due to a weak polyadenylation signal in the L1-2 sense strand. Contrarily, L1-1 in the antisense strand was truncated from bases 1 to 5,637 and did not contain the antisense strand polyadenylation signal (5′-TTTATT-3′) spanning bases 5,576–5,581. The red fluorescence was generally consistent across all assays, except for a slight increase in assay L1-2 ($t = 2.4, df = 860.5, \text{adjusted } P = 0.2$), possibly due to weaker competition from EGFP synthesis.
Fig. 5 | Somatic L1 insertions occur in genomic regions of high functional potential. L1-1 is inserted in a 2.6-kb promoter flanking region (ENSR00000032826) that is expected to regulate the expression of nearby genes. The chromatin states are shown for a subset of human cell lines. L1-1 is inserted in a LD block, based on the common SNPs that are highly correlated ($R^2 > 0.6$) with the closest common SNP to L1-1, rs1890185 (398 bp upstream of L1-1). This LD block (gray) contains 72 SNPs significantly associated with ten diseases or disorders and 28 measurement or other traits, including 13 risk SNPs from 11 schizophrenia studies.

Discussion

WGS of bulk tissue, or of cell-type fractions from a given organ, is a direct approach to detect and characterize somatic mosaicism. However, it remains challenging to discover mosaic genome variants that are individually of low mosaicism levels. Machine-learning-based approaches can improve the detection accuracy for mosaic single-nucleotide variants and indels, but the discovery of somatic MEIs faces additional challenges in both detection (for example, mapping repetitive transposon sequences) and experimental validation (for example, PCR bias). We developed a precise analytic method for detecting somatic MEIs in deep-coverage WGS data, as well as systematic experimental steps to validate the detected insertions. We used this method to detect and then define the anatomical distribution of two somatic L1 retrotransposition events in the neurons from multiple brain regions. These events demonstrated all the hallmarks of in vivo L1 retrotranspositions, with their poly(A) tails being shorter than the average length seen in previous reports but still within the plausible range. We then showed that individual somatic L1s span both brain hemispheres and are equally widespread in glia. Thus, glia, which are roughly equal in number to neurons, are also an important cell type to consider in the tracing of neurodevelopmental lineages and assessment of the potential physiological impact of somatic retrotransposition. Additionally, we envision that RetroSom will be applied to other disease states, such as various cancers, where somatic retrotransposition events can serve as driver mutations.

Two validated L1 insertions (L1-1 and L1-2) were identified in both neurons and glia cells, but not in fibroblasts obtained from the same donors, suggesting that retrotransposition likely occurred in neuroepithelial cells at the neural plate stage, before the separation of the cerebellum, basal ganglia and cortex lineages for insertion L1-1, and later in a dorsal telencephalic neuroepithelial cell for insertion L1-2. Notably, both types of neuroepithelial cells give rise to bipotential neural stem cells (the radial glia) that develop into neurons and glia and serve as a guiding scaffold.
for their migration from the developing ventricular zones to the cortical surface, with the earlier mutation event (L1-1) producing higher mosaicism levels. Previous studies demonstrated that an engineered human L1 can retrotranspose in rat hippocampal neural stem cells, human embryonic stem cell-derived neuronal progenitor cells, and can lead to neuronal somatic mosaicism in transgenic mice. Moreover, quantitative PCR experiments suggested an increase in L1 DNA copy number in several human brain regions when compared to heart or liver genomic DNAs derived from the same individual. These data hypothetically could reflect a variety of processes, including increases in neuronal aneuploidy, increases in the generation of

% Gcont-1 n = 641
GL-1-1 n = 570
Gcont-2 n = 644
GL-1-2 n = 469

% GL1-1 n = 440
GL1-2 n = 433
GL1-2 n = 338
GL1-2 n = 423

% Gcont-1 n = 271
GL1-1 n = 215
Gcont-2 n = 234
GL1-2 n = 232
single-strand L1 cDNAs and/or increases in L1 retrotransposition18-20. Since that time, several reports suggested divergent estimates regarding the rate of somatic L1 insertions in human brain. For example, two previous sequencing studies using bulk unsorted brain samples reported hundreds of putative somatic L1 insertions at 80× Complete Genomics sequencing coverage or thousands of region using targeted 30× Illumina sequencing coverage21. However, our mixing experiment indicated that sequencing at these depths would only detect insertions with relatively higher mosaicism levels (for example, >5%); our sensitivity to detect mosaicism levels >5% was 0.67, but none were observed. Subsequent single-cell sequencing studies suggested a frequency of 10 insertions19 or ≤1 insertion per neuron22,23. While our approach did not directly measure the L1 retrotransposition rate per cell, we identified and extensively validated two somatic L1s present at ~1% mosaicism, which is consistent with other findings that somatic L1 retrotransposition is relatively rare in neuronal cells. Future technological developments and a lower cost of WGS will enable even more sensitive detection, for example, also at very low (<1%) mosaicism levels, making it possible to further refine our understanding of the frequency and anatomical distribution of somatic MEIs, such as their occurrence in fetal brain tissues with incomplete clonal proliferation, in differentiated cells with limited further proliferation and in neurodevelopment where mosaicism levels are modified by tangential migration or programmed cell death24.

Can moderate or low levels of L1 mosaicism in brain have pathological consequences? Several studies have shown that somatic single-nucleotide variants present in human brain at low tissue allele frequencies (tAFs, the fraction of chromosomes carrying an alternative allele) can drive functional anomalies25-27, such as Sturge–Weber syndrome (1–18% tAF)28, local cortical dysplasia (1.3–12% tAF)29 and hemimegalencephaly (8–40% tAF)30. The identification of two somatic L1 insertions in 0.05–14.4% of brain cells (for example, 0.025–7.2% tAF) in a single individual does not establish an etiological role in neuropsychiatric disorders such as schizophrenia. But it is noteworthy that insertion L1-1 disrupted a putative transcriptional regulatory element within CNNM2, which is located within a locus that is significantly associated with schizophrenia in large-scale genome-wide analysis31, and for which knockout studies in model systems32 suggest that it may be a schizophrenia candidate gene. Insertion L1-2 disrupted FRMD4A, a gene associated with a syndrome of microcephaly and intellectual disability33, phenotypes that are also observed in carriers of genomic copy number variants that increase the risk of schizophrenia34. Lastly, both CNNM2 and FRMD4A are intolerant to loss-of-function mutations (probability of loss-of-function intolerance scores > 0.9)35.

Each individual with a genetically complex disease such as schizophrenia has a set of common risk variants and may also have rare variants with larger individual effects on risk36. The latter could include mosaic structural variations and/or MEIs with strong functional impacts that extend beyond the mutated cells in ways that are not entirely dependent on bulk-tissue mosaicism levels. In principle, these impacts could include locally disordered neurodevelopment, induction of epileptiform activity, disruption of brain circuit activity through the widespread synaptic connections of the mutated cells or altered physiology of cell–cell contacts during epiblastic cell polarization (for example, the essential role played by the FRMD4A protein in the cell adhesion protein complex)37. Thus, it is worth keeping an open mind about whether low levels of somatic MEIs contribute to neuropsychiatric disorders, and future research on this question, using much larger datasets, will be facilitated by the cost-efficient and precise method described here.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-00767-4.

Received: 15 July 2019; Accepted: 21 November 2020; Published online: 11 January 2021

References
1. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72, 595–605 (1993).
2. Richardson, S. R. et al. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. Microbiol. Spectr. https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014 (2015).
3. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. Mol. Cell 79, 9 (2016).
4. Tubio, J. M. C. et al. Extensive transduction of nonreplicative DNA mediated by L1 retrotransposition in cancer genomes. Science 345, 1251343 (2014).
5. Evrony, G. D. et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron 85, 49–60 (2015).
6. Erwin, J. A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. 19, 1583–1591 (2016).
7. Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. & Gage, F. H. The role of transposable elements in health and diseases of the central nervous system. J. Neurosci. 33, 17577–17586 (2013).
8. Jacob-Hirsch, J. et al. Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders. Cell Res. 28, 187–203 (2018).
9. Muotri, A. R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910 (2005).
10. Evrony, G. D. et al. Heritable L1 retrotransposition in the mouse primordial germine and early embryo. Genome Res. 27, 1395–1405 (2017).
11. Sanchez-Luque, F. J. et al. LINE-1 epigenesis of epileptic reprogrammings in humans. Mol. Cell 75, 590–604 (2019).
12. Balille, J. K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537 (2011).
13. Upton, K. R. et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell 161, 228–239 (2015).
14. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 151, 483–496 (2012).
15. Evrony, G. D., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. Elife 5, 1–32 (2016).
16. Zhou, W. et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz1173 (2019).
17. Rishishwar, L., Marín-Ramírez, L. & Jordan, J. K. Benchmarking computational tools for polymorphic transposable element detection. Brief. Bioinform. 18, 908–918 (2017).
18. Keane, T. M., Wong, K. & Adams, D. J. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29, 389–390 (2013).
19. Birur, R., Kraguljac, N. V., Shelton, R. C. & Lahti, A. C. Brain structure, function, and neurochemistry in schizophrenia and bipolar disorder—a systematic review of the magnetic resonance neuroimaging literature. NPJ Schizophren. 3, 15 (2017).
20. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res. 27, 1537–1564 (2017).
21. Flasch, D. A. et al. Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication. Cell https://doi.org/10.1016/j.cell.2019.02.050 (2019).
22. Breiman, L. Random forests. Mach. Learn. 45, 5–32 (2001).
23. Skowronski, J., Fanning, T. G. & Singer, M. F. Unit-length LINE-1 transcripts in human teratocarcinoma cells. Mol. Cell. Biol. 8, 1385–1397 (1988).
24. Moran, J. V. et al. Xenon shuffling by L1 retrotransposition. Science 283, 1530–1534 (1999).
25. Bae, T. et al. Different mutational rates and mechanisms in humans at a non-coding region in the human brain. Science 359, 550–555 (2018).
26. Övchinnikov, I. et al. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. https://doi.org/10.1101/gr.194701 (2001).
27. Morris, T. A. et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. 31, 159–165 (2002).
28. McConnell, M. J. et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: the brain somatic mosaicism network. Science 356, eaal1641 (2017).
Brain Somatic Mosaicism Network

Xiaowei Zhu14, Bo Zhou14, Alexander Urban14, Christopher Walsh15, Javier Ganz15, Mollie Woodworth15, Pengpeng Li15, Rachel Rodin15, Robert Hill15, Sara Bizzotto15, Alice Lee16, Alissa D’Gama16, Alon Galor16, Craig Bohrson16, Daniel Kwon16, Doga Gulhan16, Elaine Lim16, Isidro Cortes16, Joe Luquette16, Maxwell Sherman16, Michael Coulter16, Michael Lodato16, Peter Park16, Rebeca Monroy16, Sonia Kim16, Yanmei Dou16, Andrew Chess17, Attila Jones17, Chaggai Rosenbluh17, Schahram Akbarian17, Ben Langmead18, Jeremy Thorpe18, Jonathan Pevsner18, Rob Scharpf18, Sean Cho18, Flora Vaccarino19, Liana Fasching19, Simone Tomasi19, Nenad Sestan19, Sirisha Pochareddy19, Andrew Jaffe20, Apua Paquola20, Daniel Weinberger20, Jennifer Erwin20, Joohoon Shin20, Richard Straub20, Rujuta Narurkar20, Anjene Addington21, David Panchision21, Doug Meinecke21, Geetha Senthil21, Lora Bingaman21, Tara Dutka21, Thomas Lehner21, Alexej Abyzov22, Taejeong Bae22, Laura Saucedo-Cuevas23, Tara Conniff23, Diane A. Flasch24, Trenton J. Frisbie24, Jeffrey M. Kidd24, Mandy M. Lam24, John B. Moldovan24, John V. Moran24, Kenneth Y. Kwan24, Ryan E. Mills24, Sarah Emery24, Weichen Zhou24, Yifan Wang24, Kenneth Daily25, Mette Peters25, Fred Gage26, Meiyang Wang26, Patrick Reed26, Sara Linker26, Ani Sarkar26, Aitor Serres27, David Juan27, Inna Povolotskaya27, Irene Lobon27, Manuel Solis27, Raquel Garcia27, Tomas Marques-Bonet27, Gary Mathern28, Eric Cournchesne29, Jing Gu29, Joseph Gleeson29, Laurel Ball29, Renee George29, Tiziano Pramparo29, Aakrosh Ratan30 and Mike J. McConnell30

14Stanford University, Palo Alto, CA, USA. 15Boston Children’s Hospital, Boston, MA, USA. 16Harvard University, Boston, MA, USA. 17Icahn School of Medicine at Mt. Sinai, New York, NY, USA. 18Kennedy Krieger Institute, Baltimore, MD, USA. 19Yale University, New Haven, CT, USA. 20Lieber Institute for Brain Development, Baltimore, MD, USA. 21National Institute of Mental Health, Bethesda, MD, USA. 22Mayo Clinic, Rochester, MN, USA. 23Rockefeller University, New York, NY, USA. 24University of Michigan, Ann Arbor, MI, USA. 25Sage Bionetworks, Seattle, WA, USA. 26Salk Institute for Biological Studies, La Jolla, CA, USA. 27Universitat Pompeu Fabra, Barcelona, Spain. 28University of California, Los Angeles, Los Angeles, CA, USA. 29University of California, San Diego, San Diego, CA, USA. 30University of Virginia, Charlottesville, VA, USA.
Methods

Tissue collection from six human donors. We studied six human donors in this project, including an adult donor, A1S, a fetal donor, F1 and two schizophrenia–control pairs matched as closely as possible for age, brain pH, postmortem delay to autopsy and DNA integrity number: 10011^1, 11003^1, 11004^1, 12004^1 (Supplementary Table 1 and Reporting Summary). The sample size was similar to those reported in previous studies that characterized brain somatic retrotrotransposons^1^1^-^4. For donors A1S and F1, we obtained postmortem brain tissue and heart tissue after review of the proposed procedures by the Stanford University Institutional Review Board, which determined that they did not constitute human subjects research (except because research was not performed on living human subjects). Human brain tissue and fibroblasts from the schizophrenia and control donors were obtained from the Dallas Brain Collection^5^1. The clinical diagnosis for each of the schizophrenia and control donors was matched as closely as possible for age, brain pH, postmortem delay to autopsy and DNA integrity number: 10011^1, 11003^1, 11004^1, 12004^1 (Supplementary Table 1 and Reporting Summary). The yield was quantified with a hemocytometer.

Fluorescence-activated nuclear sorting. For the initial WGS screening of the adult donors, we 0.5–1 cm^3 of cortical tissue from the STG. The neuronal and glial nuclei were extracted from the postmortem brains using methods modified from a published protocol^2. Initially, the brain tissues were dissected on a cold plate (TECA LHP-1200CAS) into ~200 mg segments. For each segment, we homogenized the tissue in 3.6 ml lysis buffer (0.3 M sucrose, 5 mM calcium chloride, 3 mM magnesium acetate, 0.1 mM EDTA, 1 mM dithiothreitol, 0.1% Triton X-100 and 10 mM Tris buffer (pH 8.0)). We then added 6.5 ml sucrose buffer (1.8 M sucrose, acetate, 1 mM magnesium acetate, 3 mM calcium acetate and 0.5 mM Tris (pH 8.0)) to the bottom of the tissue lysate, and centrifuged at 100,000 g for 2 h at 4 °C (Sorvall ultracentrifuge WX-80). The nuclei in the pellet were collected by incubation in 500 μl of ice-cold PBS for 10 min, gently resuspension and filtration through a 40-μm strainer. The nuclei were washed with a protease microscope (EVOS FL), and the yield was quantified with a hemocytometer.

The neuronal and glial nuclei were separated with fluorescence-activated nuclear sorting using a BD Aria sorter that was optimized to sort nuclei based on DAPI and PE signals (Supplementary Fig. 1^5^3). We first drew gates in forward scatter (FSC-A and FSC-W), side scatter (SSC-A and SSC-W) and DAPI channels to select for singlet nuclei. The NeuN^+^ and NeuN^−^ nuclei were then separately collected with gates in the PE and FSC-A channels: NeuN^+^ nuclei are from neurons and are larger in size and carry stronger PE signals, while NeuN^−^ nuclei are from non-neurons (glial cells) and are smaller. The purity of the sorted nuclei (quantified by reanalyzing the sorted fractions) was >99.95% in both fractions. The data were analyzed with FlowJo cell analysis software (v10.0.7.r2). A typical yield of 200 mg of brain tissue is ~2 million nuclei, with NeuN^+^ and NeuN^−^ combined. The ratio between the NeuN^+^ and NeuN^−^ fraction varies depending on the anatomical region, for example, 1.6 in the STG, 12.6 in the cerebellum and 0.24 in the putamen.

Immunopanning. Immunopanning was performed using methods modified from a published protocol^1^1. In brief, fetal cortex was harvested from the elective termination of a gestational week 18 pregnancy. Cortical tissue was chopped into fine pieces (<1 mm^3) with a no. 10 scalpel blade and then incubated in 15 U ml^−^1 papain for 60 min at 34 °C, followed by digestion in 10 mM HEPES buffer, pH 7.4. The bound cells was rinsed with PBS to wash away loosely attached contaminants. Unbound cells were transferred to the subsequent petri dish, and the dish with bound cells was rinsed with PBS to wash away loosely attached contaminants. Adherent cells were dislodged with trypsin (200 units in EBSS for 5 min at 37 °C), which was briefly inactivated with fetal bovine serum before spinning and resuspending purified cells.

Genomic DNA extraction and whole-genome sequencing. The genomic DNA from neuronal nuclei, glial nuclei and non-brain controls was extracted with the Qiaquick DNA Blood & Tissue Kit. The yield is typically ~3 μg per million cells, and all DNA passed a DNA integrity number quality threshold of 7. We prepared six separate libraries for each DNA specimen, using 200 ng of genomic DNA and the Illumina TruSeq Nano DNA Sample Preparation Kit (Macrogen). These libraries were sequenced to >30x on an Illumina HiSeq X system, with a read length of 2 × 150 bp. For comparison, we also prepared two PCR-free libraries from A1S heart and A1S neuronal nuclei, each using 1 μg of genomic DNA and the Illumina TruSeq DNA PCR-free Sample Preparation Kit.

RetroSeq pipeline. Additional public datasets. We obtained several high-quality public WGS datasets for the training and testing of RetroSeq (Supplementary Table 2) as described below.

Human Genome Structural Variation Consortium. We used WGS data from three trios studied in the Human Genome Structural Variation (HGSV) Consortium, including a lymphoblastoid cell line of a Yoruban trio (NA12877 and NA12878) and 11 offspring (NA12879, NA12880, NA12881, NA12882, NA12883, NA12884, NA12885, NA12886, NA12887, NA12888 and NA12893)^3^0. All members were sequenced to an average depth of 50x (dbGaP accession: phs001224). In addition, NA12877 and NA12878 were sequenced to an average depth of 200x (ENA accession: PRJEB8326). The sequencing was carried out in PCR-free libraries on an Illumina HiSeq 2000 system, with a read length of 2 × 101 bp.

Neuropsychiatric Disease Genomics Center (NDGC). The Human Genomics Kimera Trio (HG00731, HG00732 and HG00733) and a southern Han Chinese trio (HG00012, HG00013 and HG00014)^1^4. Each cell line was sequenced with PCR-free libraries to an average depth of >30x (ftp://ftp.genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_wf_discovery/data/).

Cloned sequence datasets. The clone sequencing datasets 316 and 320 were downloaded from the National Institutes of Health (NIH) National Institute of Mental Health (NIMH) Data Archive (https://data-archive.nimh.nih.gov/) under collection ID no. 2330 and https://doi.org/10.15154/1410419 (ref. 17). Both datasets include WGS of cell clones expanded from individual neural stem cells. Dataset 316 has 25 clones amplified with multiple displacement amplification (316 WGA; n = 5), along with 71 other clones and bulk DNA from the frontal lobe and spleen (361 noWGA; n = 10); dataset 320 contains 50 clones plus bulk DNA from the basal ganglia, frontal lobe and spleen (320; n = 53).

Brain Mosaic Network Consortium common brain. We also obtained the sequencing data of the common brain tissue studied by the Brain Mosaic Network (BSMN) Consortium. The data include >200Kx WGS of the bulk brain tissue and fibroblasts.

Sequence alignment and candidate supporting reads. Raw sequencing reads from the six human donors, as well as from the public datasets, were all aligned to the human reference genome CHROM1D with the BWA-MEM aligner (BWA, v0.7.12; ‘-mem’ – t = 6 – B = 4 – O = 6 – E = 1 – M = -R) and then post-processed on alternative contigs/decov/HLA genes (bwa-postalt.js)^29. The alignment was further cleaned by removing secondary alignment, supplementary alignment and PCR duplicates. We used a modified RetroSeq pipeline (–discover –align –srmode clean) to extract candidate supporting reads with >95% identity matching the consensus sequences of L1s or Alu elements, including AluYa5, AluYa5a, AluYb8, AluYb9, AluYc1 and AluYyk13 (ref. 23). We inferred MEIs by integrating two types of supporting reads: SRs, which capture the MEI insertion point such that part of the read maps to the ME consensus sequence and the other part to the unique flanking sequence reference sequence at the new genomic location; and PE reads, where one read maps to the ME consensus (ME end) and the other to the unique flanking sequence (anchor end). The two PE supporting reads are not properly paired because the ME end is usually mapped to a distant reference ME and the sequence between the two paired reads is unknown but has a known size range. Thus, PE supporting reads help to localize the ME without giving information regarding the exact breakpoints. The SR supporting reads, on the other hand, provide breakpoint sequences but are not always available when the insertion is found in a minority of cells.

The SR supporting read has one chimeric read mapped to both the flanking sequence and the ME sequence and often contains two or three base pairs of the flanking sequence for correct mapping. Thus, the correct placement of a chimeric read requires the mate read to be properly paired. However, the BWA-MEM algorithm sometimes assigns an incorrect primary alignment location for the chimeric read even when it is properly paired with its mate. However, BWA assigns two alignments for each chimeric read: a primary alignment based on the longer segment and a supplementary alignment based on the shorter segment. When a chimeric read covers a ME I, either segment can be the primary sequence and properly paired with the mate, while the other segment will be in the ME sequence and usually mapped to a distant reference ME. When the ME segment is >50% of the chimeric read in an SR supporting read, the chimeric read is mapped to a location not properly paired with its mate in the primary alignment.
As a result, the supporting read will be reported as PE instead of SR, and the insertion junction information is lost.

To optimize the discovery of SR supporting reads, we scanned the supplementary alignment tag of all the PE supporting reads for chimeric alignments. If the position of the shorter segment could be properly paired with the anchor end, and the longer fragment could be mapped to an ME sequence, we converted the PE supporting reads to SR. Furthermore, we separately analyzed a group of PE supporting reads with an SR anchor end: the chimeric anchor end also provides additional information about the MEI junction. We ignored the PE supporting reads when <50% of their anchor ends were mapped to the flanking sequence, to avoid potential mapping errors.

We excluded supporting reads of poor quality, including those characterized by: (1) genomic regions of highly repetitive sequences, including centromeric repeats, telomeric repeats, large segmental duplications, reference genome gaps or those without a unique sequence. We first built separate random subsamples of supporting reads with low sequencing complexity (score <1) using the SEG algorithm5; or (3) outlier sequencing depth within 500 bp upstream and downstream to the insertion (more than three standard deviations away from the mean). The sequencing depth for sex chromosomes was evaluated separately. The masked reference sequence was 23.6% for L1 insertions in the positive strand, 23.7% for L1 insertions in the negative strand, 21.0% for Alu insertions in the positive strand and 21.1% for Alu insertions in the negative strand.

Simulating the putatively detectable mosaicism. We performed a simulation to evaluate the relative performance of detecting MEIs in the sequencing read data and the detectable mosaicism of somatic MEIs (Extended Data Fig. 1a). In the simulation, we assumed that (1) sequencing depth was 50x; (2) sequencing reads were 2x 150 bp in length and the fragment length (including read 1, read 2 and the insert in between) followed a normal distribution: \(N(600, 100); (3) the MEI was from 4,500 bp to 5,500 bp on a DNA segment of 10 kb in length; (4) the MEI had flanking sequences; and (5) the MEI was homozygous in the simulation model. We simulated that the sequencing fragment was shorter than the MEI and thus could not span both upstream and downstream junctions: (7) any reads that crossed the MEI junction with >30 bp overlapping with the MEI consensus and more than half of the read length (75bp) overlapping with the flanking sequences could be used as supporting reads (that is, the flanking sequence could be uniquely mapped); (8) there were no SR supporting reads from the MEI junction around the poly(A) tail because the poly(A) tail may cause inaccurate mapping of the SR (Supplementary Fig. 2).

Under these assumptions, we defined the putatively detectable mosaicism as the lowest mosaicism at which ≥50% of MEIs could be detected with a certain number of supporting reads. For instance, in a hypothetical 50x WGS dataset, the 10-kb DNA fragment containing the MEI in 0.96% of cells was expected to be covered with eight read pairs, and 32% of these MEIs were detectable with one or more supporting reads in 50,000 simulations. Similarly, the putatively detectable mosaicism was 2.24% for two supporting reads, 3.72% for three supporting reads, 5.04% for four supporting reads and 6.48% for five supporting reads (Fig. 1d). The reads in the simulation model were likely to be because MEI supporting reads may have to meet additional criteria, such as unique and high-quality mapping of the anchor-end reads. The code for the simulation is available in the Supplementary Software.

Model training. We built the RetroSom model to classify each supporting read identified in the 11 offspring from the platinum pedigree as either a true or false MEI (Extended Data Fig. 1b). For all members in the pedigree, we first identified candidate MEIs with ≥2 support reads after excluding reference MEIs, regions of highly repetitive sequences, low sequencing complexity or outlying read depth. Notably, we also separated the supporting reads from different DNA strands and called MEIs in forward/reverse strands separately. We then labeled each candidate MEI in the 11 offspring as true or false insertions based on the inheritance pattern. True insertions were transmitted from heterozygous or homozygous insertions in the parents (NA12877/NA12878). A heterozygous MEI satisfies three conditions: (1) found in a total of 1–10 offspring, each with ≥4 supporting reads; (2) found in NA12877 or NA12878, but not both, with ≥3 found in at least one of the two grandparents from either the maternal or the paternal side, but not both sides, with >4 supporting reads; and (3) found in both grandparents on either the maternal or the paternal side, but not both sides, with >4 supporting reads. A homozygous MEI satisfies another three conditions: (1) found in all 11 offspring with ≥4 supporting reads; (2) found in NA12877 or NA12878, but not both, with >4 supporting reads; and (3) found in both grandparents on either the maternal or the paternal side, but not both sides, with >4 supporting reads. We excluded MEIs present in both parents to remove common artifacts and evolutionarily ancient insertions. As expected, the occurrence of true MEIs in offspring followed a binomial distribution (Extended Data Fig. 1c).

The false insertions, on the other hand, are the ones found in the offspring but absent in both parents. There were substantial numbers of false insertions at a low cutoff of supporting reads (Fig. 1d). In the false dataset for training, we only kept low-confidence MEIs (<3 supporting reads) that were absent in both parents to exclude true de novo germine insertions in the offspring.

We built a data matrix with ‘positive’ supporting reads from true MEIs and ‘negative’ supporting reads from false MEIs; each read was characterized by a list of sequencing features (Supplementary Table 3). We followed two rules for selecting the features: (1) they should help to distinguish true retrotransposition of young active transposons from noise created from old and inactive ones, and (2) they should not cause any bias due to the limited scope of our training dataset.

Based on the first rule, we selected features that are known for the active subfamily of L1Hs element (for example, sequence identity to L1Hs consensus, ACA/G and G alleles in the 3' end) and TPRT retrotransposition model (for example, 5'-TTTT/AA-3' EN motif and no transduction for Alu). Based on the second rule, we excluded biasing features such as the number of supporting reads (limiting the sensitivity for low mosaicism insertions), features specific to individual elements (for example, unique SNPs/indels, unlikely to be shared by other families), features specific to sequencing conditions (to preserve generalizability) or chromosomal location—new retrotranspositions are believed to occur in random positions, so any positional bias in true-positive MEIs here should be due to selection and thus not relevant to somatic MEIs.

When applying the null distribution to determine numbers of supporting reads in a random forest model, we partitioned L1 PE reads into eight subgroups, with reads mapped to different segments of the L1 consensus (Extended Data Fig. 1d); L1 SR reads into two subgroups, including the original SR reads and the ones converted from PE reads; and Alu PE reads into two subgroups, including the ones with and without SR anchor ends.

To make new predictions, one candidate L1 PE supporting read may be categorized into several subgroups and therefore have multiple probability scores. RetroSom reports the probability based on the sub-model with the best accuracy, in the following order: (1) RFI.1, (2) RFI.4, (3) RFI.6, (4) RFI.2, (5) RFI.5, (6) RFI.7, (7) RFI.6 and (8) RFI.3. The order is based on the overall accuracy of each model in the training dataset (Extended Data Fig. 1e). Most sub-models produced highly similar predictions, and the ranking had little impact on the overall prediction. We chose the default probability score cutoff (0.5) for classifying new supporting reads as true MEI insertions. The scripts for the modeling are available in the Supplementary Software.

Evaluation training data with 11x cross-validation. The performance of RetroSom was first evaluated with 11x cross-validation. Each of the 11 offspring was selected as the test dataset once, while the data from the remaining 10 offspring were used for modeling. For comparison, we also built a logistic regression model (LogR), a Lasso regression model (Lasso), a Ridge regression model (Ridge) and a naïve Bayes model. The machine learning was performed in R (3.5.0); logistic regression (with and without regularization) with the 'glmnet' package (v2.0-16), random forest used the 'randomForest' package (v4.6-14) and naïve Bayes used the 'e1071' package (v1.6-8)59.

We evaluated the models using six metrics: (1) accuracy = (true positive + true negative) / (false positive + false negative + true positive + true negative); (2) precision = true positive / (true positive + false positive); (3) sensitivity = true positive / (true positive + false negative); (4) specificity = true negative / (true negative + false positive), (5) area under receiver operating characteristic curve and (6) area under precision-recall curve. The area under receiver operating characteristic curve and area under precision-recall curve were calculated using the 'PRROC' package (v1.3.1; Extended Data Fig. 1f).

Evaluation in fetal brain clonal expansion. We evaluated RetroSom in two public clone sequencing datasets, 316 and 320, created by culturing individual neural cells from fetal brains and sequencing genomic DNA from each clone25. Dataset 316 contains 50 clones and bulk DNA from two brain regions and one non-brain tissue. In addition to being single-cell clones, these datasets differed from the Platinum dataset 320 contains 50 clones and bulk DNA from two brain regions and one non-brain tissue. Based on the first rule, we selected features that are known for the active subfamily of L1Hs element (for example, sequence identity to L1Hs consensus, ACA/G and G alleles in the 3' end) and TPRT retrotransposition model (for example, 5'-TTTT/AA-3' EN motif and no transduction for Alu). Based on the second rule, we excluded biasing features such as the number of supporting reads (limiting the sensitivity for low mosaicism insertions), features specific to individual elements (for example, unique SNPs/indels, unlikely to be shared by other families), features specific to sequencing conditions (to preserve generalizability) or chromosomal location—new retrotranspositions are believed to occur in random positions, so any positional bias in true-positive MEIs here should be due to selection and thus not relevant to somatic MEIs.

In the fetal brain dataset, the machine learning was performed in R (3.5.0); logistic regression (with and without regularization) with the 'glmnet' package (v2.0-16), random forest used the 'randomForest' package (v4.6-14) and naïve Bayes used the 'e1071' package (v1.6-8)59.

We evaluated the models using six metrics: (1) accuracy = (true positive + true negative) / (false positive + false negative + true positive + true negative); (2) precision = true positive / (true positive + false positive); (3) sensitivity = true positive / (true positive + false negative); (4) specificity = true negative / (true negative + false positive), (5) area under receiver operating characteristic curve and (6) area under precision-recall curve. The area under receiver operating characteristic curve and area under precision-recall curve were calculated using the 'PRROC' package (v1.3.1; Extended Data Fig. 1f).
selected based on their presence in all 20 libraries, including 18 TruSeq Nano (three cell fractions) and 2 PCR-free sequencing datasets. True MEIs were selected as the insertions that were highly supported in most of the libraries (more than four supporting reads in >80% libraries), while false MEIs were selected as the insertions that were missing or poorly supported in most of the libraries (less than three supporting reads in >80% libraries).

**Evaluation in mixed DNA with different frequencies.** To evaluate the performance of RetroSom in detecting MEIs with low levels of mosaicism, we developed a sequencing experiment to use genomic DNA mixed at various frequencies to simulate real mosaic MEIs. We first spiked six unrelated genomic DNA in NA12878 DNA at a gradient of concentrations, including (1) A15 heart at 0.04%, (2) NA19240 at 0.2%, (3) HG00733 at 1%, (4) HG00514 at 1%, (5) BSMN common brain at 5% and (6) NA12878 at 25%. The mixed DNA was meant to simulate a specimen containing different MEIs of different frequencies, while pure NA12878 was meant to simulate a control specimen without any somatic MEIs. The DNA spiked in was chosen based on the following three criteria. First, the chosen DNA was either sequenced deeply (>200x) by our group (A15 and BSMN brain) or included as the chimer in trios chosen by the HGSV (NA19240, HG00733 and HG00514) or Platinum Genomes (NA12878 and NA12878). Based on the existing sequencing data, we created a high-confidence catalog of MEIs that are unique to each DNA. Notably, homozygous MEIs are presented in the mixed DNA at a frequency twice as high as the heterozygous MEIs. To better simulate real somatic MEIs that are almost certainly heterozygous when occurring, we only considered heterozygous MEIs in each of the spiked genomes. Second, we chose DNA of distinct ancestries to minimize the number of unique MEIs at each mosaic level. Most of the genomic DNA has a low level of heterozygous L1 insertions that are not shared with anyone else (between 11 and 32), except for the African sample NA19240, which has 77 unique L1. We speculated that the detection sensitivity of our 200x bulk sequencing was between 0.2% and 1%, and decided to have more unique L1s spiked at these two ratios. As a result, we spiked NA19240 at 0.2% and both HG00733 and HG00514 at 1%. Thirdly, NA12878 was chosen as the backbone in the mixing experiment because it is from a homogeneous cell line and is one of the most well-studied genomes.

The unique heterozygous MEIs in each of the spiked DNA samples were defined as: $\text{Unique}_\text{MEI} = \{ \text{MEI}_i \mid j \neq i \} \cup \{ \text{MEI}_j \mid i \text{ is one of the six DNA spiked at ratio } 0.04\% \text{ to } 25\% \text{ and } j \text{ is one of six spiked DNA or NA12878 (} j = 7\text{)} \}$ For both of the mixed DNA (named ‘mix’) and pure NA12878 (named control), we made six separate libraries (TruSeq Nano) and sequenced each library to an average depth of 30–40x (total = 200x). We applied RetroSom to call somatic MEIs that were found in the mixed DNA but not in the NA12878 control. The false positives and true positives were then defined as:

$$
\text{MEI}_{\text{false, positive}} = \text{MEI}_{\text{mix}} - \text{MEI}_{\text{control}} - \bigcup_{i=1}^{6} \text{MEI}_i
$$

$$
\text{MEI}_{\text{true, positive}} = \left( \text{MEI}_{\text{mix}} - \text{MEI}_{\text{control}} \right) \cap \left( \text{Unique}_\text{MEI} \right)
$$

$\text{MEI}_{\text{false, positive}}$ is the set of MEIs called from the 200x sequencing of mixed DNA; $\text{MEI}_{\text{true, positive}}$ is the set of MEIs called from the 200x sequencing of NA12878 control; and $\text{MEI}_i$ is one of the MEIs spiked from 0.04% to 25%. To evaluate the performance at different read depths, we downsampled the sequencing data (mix and control) to 50x and 100x sequencing with Picard (DownsampleSam; v2.17.3). We also mixed raw reads from previous sequencing datasets of each component at the same frequencies to create an in silico mixing dataset of 200x and combined it with the mix sequencing data to a final depth of 400x. The sources included our own sequencing (A15, HGVS dataset (HG00733, HG00514 and NA19238), BSMN common brain and the 200x Platinum Genomes dataset (NA12878)). The 400x control data were created by combining the 200x NA12878 WGS in the Platinum Genomes and the control sequencing data. Notably, we did not re-use the training data for testing at 400x. Notably, we did not re-use the training data for testing at 400x.

**Manual curation to remove false MEIs.** To select a set of MEIs for experimental validation, we adopted a series of manual inspections to further eliminate likely false positives (Extended Data Fig. 4). We first examined the neighboring region of the candidate MEIs, removing putative insertions likely caused by structural variation or in regions with poor mapping quality (using the integrated genomics viewer IGV)\(^\text{7}\). We also removed somatic MEIs present in datasets from other donors, likely occurring in regions prone to sequencing and mapping artifacts. We then used the visualization tool RetroVis to plot each insertion and its supporting reads, allowing for a rapid screening of multiple candidate MEIs. Finally, we compared the sequences of the supporting reads to remove false insertions characterized by unexpected transduction, conflicting positions between support or low homology in the ME ends mapped to the same location. The majority of the putative somatic MEIs were filtered during the manual curation, and the exact filters used are listed in Supplementary Table 4.

**Supporting reads for L1-1 and L1-2.** L1-1 was discovered with two supporting reads and L1-2 with three supporting reads. The reads were trimmed for sequencing adaptors, low-quality ends and flanking N bases (cutadapt: --a AGATCGGAAGAGC --AAGATCGGAAGAGC --trim --n --m 20 --v 1.8.1); L1-end, read that maps to LHS consensus sequence; anchor_end, read that maps to the flanking sequence; underline, mismatching bases outside of poly-A tracts.

To post-processing of putative somatic MEIs in RetroVis package to visualize the supporting reads. RetroSom includes a visualization tool, RetroVis, that systematically visually analyses the supporting reads for each putative MEI with clear annotations for the insertion position, orientation and other vital information (Extended Data Fig. 4a). Traditional genome browsers have issues with displaying the positions of both the anchor ends in the flanking sequences and the ME ends in the L1/1AU consensus. In addition, supporting reads for somatic MEIs are few in number and usually only captured by other sequencing reads nearby. The scripts for RetroVis are available in the Supplementary Software.

In RetroVis, we annotated the human reference genome around the insertion junction as a black line on the top and the ME consensus on the bottom. The segment coordinates are labeled above the lines, and a short vertical line marks every 20 bp. In the reverse strand, both the line are the PE and supporting inserts are labeled. Each PE supporting read is represented by a pair of arrows: a blue arrow and a red (or purple) arrow connected by a dashed line. The blue arrow represents the read that mapped to flanking human genome sequences, and its location is based on the human reference on the top. The red (or purple) arrow represents the read that mapped to the ME consensus, and its location is based on the ME consensus on the bottom. A red arrow indicates the ME is inserted in the forward strand, while a purple arrow indicates the insertion is in the reverse strand. For the SR supporting read, the chimeric read that covered the insertion junction is plotted as a blue arrow connected to an empty rectangle. The blue arrow represents the read segment that mapped to the flanking sequences, while the empty rectangle represents the ME segment, the alignment of which is indicated by a red/purple arrow below. This visualization provides a convenient way to manually check any MEIs, especially when picking candidates for experimental validation.
ddPCR assays were prepared using a published protocol. The primer and probe sequences characterized by overlap extension PCR (Supplementary Fig. 3). For sequences outside of poly(A) tracts, 1,245 of 1,250 bases were perfect matches (identity = 99.6%). The five mismatching bases, underlined in L1-2_support1_read1, could not be mapped to L1-1 or the flanking end of L1-1_support1_read1 following the poly(A) tail (ACCCAAAAAAATC′). More mismatches were found in the L1 poly(A) tracts. For instance, the 3′ end of L1-1_support1_read1 could not be mapped to other locations in the human reference genome with a high-confidence global alignment, and thus it is unlikely to be a PCR chimera connecting a reference L1 sequence to the flanking sequence as indicated in L1-1_support1_read1. The low GC percentage, low sequence complexity and poor sequencing qualities are all indicators of sequencing errors.

The reactions for L1 insertions were incubated as follows: 95°C for 10 min; 94°C for 30 s (50 cycles) 59°C for 1 min (30 cycles); 98°C for 10 min. The cutoffs separating the positive and negative droplets were chosen based on the negative and positive controls, and the levels of mosaicism were quantified using QuantaSoft Analysis Pro Software (v1.0; Bio-Rad). The target allele frequency was calculated from the number of positive droplets, based on the method described by Zhou et al. Under the assumption that somatic MEIs are heterozygous, their levels of mosaicism were calculated to be twice the allele frequency.

*Nested PCR.* We used two rounds of PCR to sequence the upstream and downstream junctions of the somatic MEIs (Extended Data Fig. 5a). In the first PCR, we used primers on the flanking sequences surrounding the MEI and 60 ng genomic DNA extracted from the right STG neurons. The pre-integration allele was present in >99% of the cells and produced a strong band consistent with the coordinates in the human reference genome. The MEI-containing allele was expected to produce a larger band. Sequencing from this band was usually invisible on gel electrophoresis because of the amplification bias towards shorter and higher-frequency products (Supplementary Note 5). Nevertheless, we purified the DNA above the visible band from the first PCR, from a region that was 270–870 bp above for L1-1 or 260–610 bp above for L1-2 (Zymocut Gel DNA recovery kit; Zymo Research, D4007). In the second round of PCR, half of the purified DNA was used to amplify the upstream junction using a primer in the upstream flanking sequence and a primer in the ME sequence. The other half was used to amplify the downstream junction using a primer in the downstream flanking sequence and a primer in the ME sequence. The nested PCR produced clean bands of expected size covering the upstream and downstream junctions, which were then analyzed with Sanger sequencing (sequencher). Combining the junction sequences, we aligned them to the target MEI using the Human Genome Build 37 (hg19). Sanger sequencing confirmed that the junctions were created by the same MEI insertion, and therefore matched the expected MEI junction, TSD, endonuclease cutting sites, inserted ME sequences and the microhomology between the ME sequence and the target site sequence. The insertion was 5′ truncated. If there was a homology between the ME poly(A) tail and the TSD, we arbitrarily included the homologous region as part of the TSD sequence. We defined 5′ microhomology by the allowance of up to one mismatching base between the L1Hs and the target site sequence. All PCR reactions were incubated in a volume of 40 μl, containing 20 μl Phusion green Hotstart II HF PCR master mix (2X; Thermo Fisher), 0.9 μM of the primers and the relevant template DNA. The primer sequences are available in Supplementary Table 6. The reactions were incubated as follows: 94°C for 2 min; 94°C for 30 s (30 cycles); 55°C for (L1-1) or 59°C (for L1-2) for 15 s (30 cycles); 72°C for 1 min (30 cycles); 72°C for 5 min.

**Spatial distribution of L1-1 and L1-2.** We sampled 12 additional pairs of tissues from symmetric regions in both hemispheres from the brain of donor 12004, including (1 and 2) two pairs in STG (BA22), (3) superior frontal gyrus (marked as prefrontal cortex proximal to STG; BA9), (4) inferior frontal gyrus (marked as prefrontal cortex proximal to STG; BA46), (5) motor cortex distal to STG (BA6), (6) motor cortex proximal to STG (BA46), (7) superior parietal lobe (marked as parietal cortex distal; BA7), (8) inferior parietal lobe (marked as parietal cortex proximal; BA39), (9) occipital cortex distal to STG (BA19), (10) occipital cortex proximal to STG (BA19), (11) putamen and (12) cerebellum (Extended Data Fig. 8a). We separated the neurons and glial nuclei with fluorescence-activated cell sorting and used ddPCR to test for the presence and mosaicism of L1-1 and L1-2 in the genomic DNA of neurons and glia, respectively. Each DNA was tested in four technical replicate experiments using 30 ng of genomic DNA. The levels of mosaicism were calculated as twice the allele frequency, and we set the ddPCR detection threshold at >0.05% mosaicism (>1 positive L1 junction droplet per replicate). The correlation between the mosaicism levels in neurons and in glia is shown in Extended Data Fig. 8b.
Bam FRMD4A is in the reverse strand of the insert size. In addition, we incubated the 5 ml LB culture at 37 °C overnight with 7.75 μl master mix. L1-1 was mixed in a 11.7 μl reaction containing 61.25 ng L1-1 plasmid DNA and 5 μl Opti-MEM medium. The reactions were incubated as follows: 95°C for 2 min
94°C for 45 s (30 cycles)
57°C for (L1-1) 1 min (for L1-2) for 30 s (30 cycles)
72°C for 2 min (30 cycles)
72°C for 7 min

Cloning into plasmid pGint. The L1-1 and L1-2, as well as the two control DNA, were digested using BamHI and Apal enzymes (NEB nos. R3136S and R01145, respectively). We first incubated 1 μl purified DNA with 1 μl Apal enzyme and 3 μl NBT CutSmart (10x) buffer in a total volume of 29 μl for 2h at 25°C. Then we added 1 μl BamHI-HF enzyme and incubated at 37°C overnight. The reaction was stopped by adding 6 μl pure loading dye. The digested DNA was gel purified (Zymoclean Gel DNA recovery kit, D4007) and eluted in 10 μl water. The DNA concentration was quantified with Qubit (LifeTech, Q33216).

The DNA fragments were ligated to the pGint plasmid using the Instant Sticky-end Ligase (2x) master mix (NEB no. M0370S) at threefold (insert sequence with Sanger sequencing. The validated clones were named GL1-1 (1,123-bp insertion with L1-1 and flank), GL1-2 (686-bp insertion with L1-2 and flank), Gcont-1 (691-bp insertion with flanking sequence for L1-1) and Gcont-2 (240-bp insert with flanking sequence for L1-2).

Transient transfection of reporter plasmids into HeLa cells. The four plasmids, Gcont-1, Gcont-1, Gcont-2 and GL1-2, were transfected into HeLa S3 cells with lipofectamine 3000 reagent in two separate experiments: (1) dual transfection together with a red fluorescence protein reporter (RFP) ‘Rint’ in five wells per reporter (Fig. 6c) and (2) single transfection without a single reporter (Fig. 6c). For the dual-transfection experiment, we seeded HeLa cells on a 24-well plate (~70% confluence, 50,000 cells per well). On the next day, we prepared the lipofectamine mixture containing 33.75 μl lipofectamine 3000 (Thermo Fisher) and 562.5 μl Opti-MEM medium (Thermo Fisher). We then prepared a plasmid DNA mixture for each reporter as follows: (1) 4.36 ng Gcont-1 plasmid (375 ng μl−1) with 1.95 μl Rint plasmid (909 ng μl−1), 145.4 μl Opti-MEM medium and 6 μl P30000 reagent; (2) 6.64 μl GL1-1 plasmid (266 ng μl−1) with 1.95 μl Rint plasmid, 147.8 μl Opti-MEM medium and 6 μl P30000 reagent; (3) 1.01 μl Gcont-2 plasmid (780 ng μl−1)) with 56.25 μl Opti-MEM medium and 1.125 μl P3000 reagent; (4) 1.10 μl GL1-2 plasmid (1,480 ng μl−1) with 1.95 μl Rint plasmid, 150 μl Opti-MEM medium and 6 μl P3000 reagent. The same number of copies of plasmids was used in each mixture, as the amount was calculated based on the plasmid size as follows: Gcont-1 = 5,410 bp, GL1-1 = 5,482 bp, Gcont-2 =2,4959 bp, GL1-2 = 5,405 bp and Rint = 5,816 bp. Per each plasmid, we mixed 133.75 μl lipofectamine mixture with 133.75 μl plasmid mixture, incubated the mixture at room temperature for 15 min and applied 50 μl to each of the five wells. The order of each transporter assay was shuffled and kept hidden until the fluorescence was quantified by a different experimenter to allow for a blind experiment.

A similar protocol was used in the single-transfection experiment, except for the plasmid mixtures (preparing the plasmid mixture)7: (1) 0.775 μl Gcont-1 plasmid (780 ng μl−1) with 56.25 μl Opti-MEM medium and 1.125 μl P3000 reagent; (2) 0.745 μl GL1-1 plasmid (890 ng μl−1) with 56.25 μl Opti-MEM medium and 1.125 μl P3000 reagent; (3) 0.380 μl Gcont-2 plasmid (1,480 ng μl−1) with 56.25 μl Opti-MEM medium and 1.125 μl P3000 reagent and (4) 0.416 μl GL1-2 plasmid (1,480 ng μl−1) with 56.25 μl Opti-MEM medium and 1.125 μl P3000 reagent.

Fluorescence quantification. After incubating HeLa cells with the transfection mixtures for 23h, we captured images in GFP, RFP and bright-field channels (Leica DMI 3000B) in each well on the top-center, bottom-left and bottom-right sections (Fig. 6a, b). We counted the number of cells (Extended Data Fig. 10). The GFP and RFP images were taken with an exposure time of 200 ms and analog gain of 9, and the bright-field images were taken with an exposure time of 40 ms and analog gain of 2. This process took ~2h in the dual-transfection experiment, so we followed a special order of measurement to avoid time-related bias (Fig. 6b). We also confirmed, in a separate pilot experiment, the absence of bleed-through interference between the GFP and RFP channels.

On each image, we labeled all cells with visible fluorescence signals (green or red) with a region of interest (ROI) marker that were adjusted to fit the cell shape, as well as five blank regions (top left, top right, center, bottom left and bottom right), to measure the background fluorescence (Leica Application Suite 300 build 8134; Fig. 6d and Extended Data Fig. 10). We used RFI = Background to represent the signal strength of each cell, where RFI represents the mean intensity value of pixels in the ROI, and Background represents the mean intensity value in all five blank regions. We excluded dead/broken cells and image artifacts by referring to the bright-field image. The number of plasmids transfected into each cell is highly variable but the impact from each reporter can be evaluated after averaging a large number of cells. From two independent experiments and seven wells per plasmid, we quantified 912 cells for GCont-1, 785 cells for GL1-1, 877 cells for GL1-2 before the plasmid labels were revealed for statistical analysis.

Estimation of poly(A) tail sizes. We evaluated the length of poly(A) tails in 24 GL1-1 clones and 24 GL2-1 clones using Sanger sequencing (Extended Data Fig. 8c). The variable poly(A) tails are likely caused by polymerase slippage around low-complexity sequences, leading to both longer and shorter poly(A) sizes7. We chose the size supported by the highest number of clones as the estimates for the number of poly(A) sizes required PCR amplification from the tissue DNA and may have introduced biases towards shorter products and templates with higher mosaicism.

PCR bias in co-amplification of the pre- and post-integration sites. To illustrate the PCR bias when amplifying the pre- and post-integration sites together, we tested amplification on a concentration gradient of a known L1 template extracted from the reporter plasmid GL1-1, including 248 bp upstream, 449 bp L1-1 and 429 bp downstream sequence. We added 1 × 10^{-6} ng, 1.43 × 10^{-6} ng, 2.04 × 10^{-6} ng, 2.92 × 10^{-6} ng and 4.16 × 10^{-6} ng of the L1-1 template (1,126 bp) to 22.8 ng NAI2878 genomic DNA to make allele frequencies of L1-1 at 92.4%, 64.7%, 20.7%, 3.59% and 0.01% respectively. We then tested PCR amplification with external primers in the flanking sequences, using PhusionTaq or DreamTaq polymerases, and 30 or 60 PCR cycles (Extended Data Fig. 5c). The PhusionTaq PCR reactions were incubated in a volume of 20 μl, containing 10 μl Phusion green Hotstart HF PCR master mix (2x; Thermo Fisher), 0.9 μM of the primers and the relevant template DNA. The primer and L1-1 template sequences are available in Supplementary Table 6. The reactions were incubated as follows: 94°C for 2 min
94°C for 45 s (30 cycles)
55°C for 30 s (30 cycles)
57°C for 2 min (30 cycles)
72°C for 4 min

We examined the PCR product on gel electrophoresis to check for the correct insert size, and additionally, we incubated the 5 μl LB culture at 37 °C overnight with shaking. We extracted the plasmid using Miniprep (Qiagen, 27106) and verified the insertion sequence with Sanger sequencing. The validated clones were named as follows: GL1-1 (1,123-bp insertion with L1-1 and flank), GL1-2 (686-bp insertion with L1-2 and flank), Gcont-1 (691-bp insertion with flanking sequence for L1-1) and Gcont-2 (240-bp insert with flanking sequence for L1-2).
Verification of the L1 post-integration site with droplet-based full-length PCR. For L1-1, we prepared eight droplet-based full-length PCR reactions from the genomic DNA of glia in two brain regions—LSTG2 and LOP—with NA12878 genomic DNA as negative controls and the L1-1 template in plasmid GL1-1 as positive controls (Extended Data Fig. 6d). Each reaction was incubated in 20 μl containing 30 ng genomic DNA, 0.9 μM primers in the flanking sequences (P1 and P2), 0.25 μM FAM probe (in L1) and 10 μl ddPCR supermix for probes (no dUTP; Extended Data Fig. 6e). Sequences for the primers and probes are listed in Supplementary Table 6. The reactions were incubated in a condition adapted for long amplicons:

- 95°C for 10 min
- 94°C for 30s (30 or 60 cycles)
- 55°C for 30s (30 or 60 cycles)
- 72°C for 1 min (30 or 60 cycles)
- 72°C for 10 min

Statistical analysis. We used Welch’s two-sided t-test to calculate the statistical significance of the mosaicism difference in various fractions (Fig. 3b,c,e). The correlation of L1 mosaicism levels between neurons and glia in different anatomical regions was evaluated by rank-based Spearman ρ statistic (Extended Data Fig. 8b).

To evaluate the level of fluorescence in the transfection experiments (Fig. 6f–h), we performed a log transformation of the fluorescence intensities to analyze and report the fluorescence levels. A dummy variable was added to all fluorescence values to remove 0 and negative values, and the log transformation was used to transform the fluorescence values to approximately conform to normality, but this was not formally tested.

We performed ten statistical tests comparing various groups in the transfection experiments and adjusted the P value using the Bonferroni correction: adjusted P = value

A possible explanation for the lower fluorescence level in L1 reporters compared to that in controls is slower transcription due to larger insert_length (Fig. 6h). However, our data suggest that the difference in the tested range of insert_length (240 to 1,123 bp) is unlikely to be the only contributing factor to the difference between L1 and control reporters. The fluorescence in GLcont-1 (686 bp) was similar to that in GLcont-2 (240 bp; adjusted P = 1) but significantly stronger than that in GL1-1 (691 bp; adjusted P = 2.6 × 10⁻⁷). In addition, the fluorescence in GL1-1 was stronger than in GL1-2 (adjusted P = 3 × 10⁻⁷), despite the larger insert_length (1,123 bp versus 691 bp).

To further evaluate the impact of insert size, we built a linear regression model to fit the GEP fluorescence for all four plasmids: log(GEP) = L1 + log(insert_length), where L1 is a binary variable indicating whether the plasmid has an L1 insertion (L1 = 1) or is a control L1 = 0, and log(insert_length) = 691 for GLcont-1, 1,123 for GL1-1, 240 for GLcont-2 and 686 for GL1-2. In this linear model, the insert_length did not affect the fluorescence intensity significantly (adjusted P = 0.25, coefficient = 0.10), while L1 was negatively correlated with the fluorescence intensity (adjusted P = 4 × 10⁻⁷, coefficient = −0.485).

References
50. Stan, A. D. et al. Magnetic resonance spectroscopy and tissue protein concentrations together suggest lower glutamate signaling in dentate gyrus in schizophrenia. Mol. Psychiatry 20, 433–439 (2015).
51. Matevossian, A. & Akbarian, S. Neuronal nuclei isolation from human postmortem brain tissue. J. Vis. Exp. https://doi.org/10.3791/914 (2008).
52. Kozlenkov, A. et al. A unique role for DNA (hydroxym)ethylation in epigenetic regulation of human inhibitory neurons. Sci. Adv. https://doi.org/10.1126/sciadv.aau6190 (2018).
53. Julius, M. H., Masuda, T. & Herzenberg, L. A. Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. Proc. Natl Acad. Sci. USA 69, 1934–1938 (1972).
54. Zhang, Y. et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. Neuron 89, 37–53 (2016).
55. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. Nat. Genet. 52, 76–81 (2015).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
57. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16, 418–420 (2000).
58. Wootton, J. C. Non-global domains in protein sequences: automated segmentation using complexity measures. Comput. Chem. 18, 269–285 (1994).
59. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22 (2010).
60. Liu, A. & Wiener, M. Classification and regression by randomForest. R News 2, 18–22 (2002).
61. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77 (2011).
62. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
63. Zhou, R. et al. Detection and quantification of mosaic genomic DNA variation in primary somatic tissues using ddPCR: analysis of mosaic transposable-element insertions, copy-number variants and single-nucleotide variants. Methods Mol. Biol. 1768, 173–190 (2018).
64. Szak, S. T. et al. Molecular archeology of L1 insertions in the human genome. Genome Biol. 3, research0052.1 (2002).
65. Beckman, K. L. & Pease, L. R. Gene splicing and mutagenesis by PCR-driven overlap extension. Nat. Protoc. 2, 924–932 (2007).
66. Bonano, V. I., Oltew, S. & Garcia-blanco, M. A. A protocol for imaging alternative splicing regulation in vivo using fluorescence reporters in transgenic mice. Nat. Protoc. 2, 2166–2181 (2007).
67. Shinde, D., Lai, Y., Sun, F. & Arnheim, N. Taq DNA polymerase slippage mutations rates measured by PCR and quasi-likelihood analysis: (CA/GT), and (A/T), microsatellites. Nucleic Acids Res. 31, 974–980 (2003).
Acknowledgements

We thank W. H. Wong, J. Chao, A. Z. Wang and N. Bosch for constructive comments on the manuscript. We thank J. E. Klemman, T. H. Hyde and D.W. from Lieber Institute for Brain Development for providing the BSMN common brain tissue and L. Fasching from Yale University for extracting the BSMN common brain DNA. This work utilized computing resources provided by the Stanford Genetics Bioinformatics Service Center. Funding: this work was supported by Eureka Grant R01MH094740 from the NIMH and the Stanford Schizophrenia Genetics Research Fund. The mixing-genome DNA sequencing and BSMN common brain sequencing data were generated as part of the BSMN Consortium and supported by: U01MH106874, U01MH106876, U01MG106882, U01MH106883, U01MH106884, U01MH106891, U01MH106891, U01MH106891, U01MH106891, U01MH106892, U01MH106893, and U01MH108896 awarded to N.S., F.M.V., F.G., C.W., P.P., J.P., A.C., J.V.M., D.W. and J.G. B.Z. is funded by the National Heart, Lung, and Blood Institute grant T32 HL110952. A.E.U. was a Tashia and John Morgridge Faculty Fellow of the Stanford Child Health Research Institute. The Urban laboratory receives funding through the Jaswa Innovator Award and from B. Blackie and W. McVor. We acknowledge helpful discussions with B. Blackie and W. McVor. Flow cytometry sorting was performed on an instrument in the Stanford shared fluorescence-activated cell sorting facility obtained under an NIH S10 Shared Instrument Grant (S10RR025518-01).

Author contributions

X.Z. coordinated the project, wrote the manuscript, and designed the model and the computational framework, with initial advice from A.F. and D.P. X.Z., B.Z. and R.P. designed and carried out the MEI validation experimental approaches. K.G., C.T., C.A.T., S.S., B.A.B. and H.V. provided the tissue samples. J.M., A.A. and F.M.V. provided the clone sequencing data. X.Z. and B.Z. generated the genome-mixing data. A.K. performed the transfection in the reporter assays and X.Z. quantified the data. L.D. advised the polygenic risk score analysis. J.V.M. contributed to the interpretation of the somatic L1 sequences. A.E.U. conceived the original idea. D.F.L. and A.E.U. supervised the project. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Competing interests

J.V.M. is an inventor on patent US6150160, is a paid consultant for Gilead Sciences, serves on the scientific advisory board of Tessera Therapeutics (also receives consultant fees and has equity options), and currently serves on the American Society of Human Genetics Board of Directors. C.A.T is or has been a deputy editor for the American Psychiatric Association; an ad hoc consultant for Astellas, Merck and Lundbeck; a council member for the Brain & Behavior Research Foundation, the National Academy of Medicine, the National Alliance on Mental Illness and a reviewer for the NIMH; she is an advisor for Karuna Therapeutics and owns its stock.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-020-00767-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-00767-4.
Correspondence and requests for materials should be addressed to A.E.U.
Peer review information Nature Neuroscience thanks Geoffrey Faulkner, Alysson Muotri and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1  | Classification of supporting reads from putative mobile element insertions.  

**a**, We simulated the relationship between the detectable mosaicism of somatic MEIs and the number of supporting reads in bulk sequencing by considering the range of coordinates for the putative supporting reads for either the upstream or downstream junction (see Fig. 1d). Blue, segment of supporting read that maps to flanking sequence; red, segment of read that maps to ME consensus; gray, the insert segment between the two paired-end reads.  

**b**, A detailed flowchart describing the framework behind RetroSom. We labeled putative supporting reads as true or false insertions based on the inheritance pattern and built a set of random forest models to classify them based on various sequencing features (see Supplementary Table 3).  

**c**, The distribution of true L1 (left) and Alu (right) insertions among 11 offspring is similar to a theoretical binomial distribution (red line). The peaks around N = 11 represent additional MEIs that are homozygous in one of the parents and transmitted to all 11 offspring.  

**d**, To avoid missing values, we categorized L1 PE supporting reads into 8 subgroups depending on their mapping locations on the L1Hs (L1 human specific) consensus sequence.  

**e**, The performance of random forest classification in all 8 L1 PE read sub-models, ranked based on their average F1 score (harmonic average of sensitivity and precision) from 11x cross validation (n = 11 tests).  

**f** and **g**, Model selection and evaluation with 11x cross validation: (**f**) precision-recall curve, (**g**) area under the precision-recall curve (AUPR, n = 11 tests). The boundaries of the boxplots indicate the 25th percentile (above) and the 75th percentile (below), the black line within the box marks the median. Whiskers above and below the box indicate the 10th and 90th percentiles.
Extended Data Fig. 2 | Benchmarking Alu insertions in independent test datasets.  

**a**, Performance in detecting germline Alu insertions from clonally expanded fetal brain cells sequencing data. Gray, clones from donor “316” sequenced with whole genome amplification (316WGA, n=10 clones); brown, the rest of the “316” datasets (316 noWGA, n=5 clones); blue, clones from donor “320” (n=52 clones). The boundaries of the boxplots indicate the 25th percentile (above) and the 75th percentile (below), the black line within the box marks the median. Whiskers above and below the box indicate the 10th and 90th percentiles.  

**b**, Performance in detecting germline Alu insertions from sequencing libraries prepared with or without PCR. Light blue/green, PCR-free libraries for sample “Heart” (light blue circle, n=1) and “Neuron” (light green triangle, n=1); Dark blue/green, PCR-based libraries for “Heart” (dark blue circle, n=6) and “Neuron” (dark green triangle, n=6). **c-e**, Performance in detecting somatic MEIs simulated by six genomic DNA samples at proportions of 0.04% to 25% with that of NA12878, at various sequencing depth (gray, 50x; brown, 100x; blue, 200x; green, 400x).
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | Discovery and experimental validation of insertion L1-3. 

**a.** We identified a somatic L1 insertion (L1-3, red arrow) in one clone, “BG clone16,” with 17 supporting reads.

**b.** L1-3 is inserted into an intron of gene EVC2. Blue, segment of supporting read that maps to the flanking sequence; red, segment of read that maps to ME consensus.

**c.** PCR (n=1 replicate) surrounding L1-3 produced a unique band in BG clone16, as well as a lower band in all tested samples, representing the product from the DNA without the insertion.

**d.** DdPCR (n=2 replicates) detects the upstream junction in 22.54% of the cells in BG clone16.

**e.** DdPCR (n=2 replicates) detects the downstream junction in 24.16% of the cells in BG clone16.

**f and g.** L1-3 is absent in 6 bulk tissues (n=4 replicates): BG ventricular zone/subventricular zone (BG VZ/SVZ), BG cortex (BG CX), FR VZ/SVZ, FR CX, occipital cortex, and spleen. The error bars represent the 95% confidence intervals of the mosaicism level in BG clone16.

**h.** The full sequence of L1-3: black, flanking sequence; red, inserted L1 sequence; purple, target site duplication; brown, mismatches to the L1Hs consensus.

**i.** Sequencing depth and reads around L1-3 junction in BG clone16. Mismatch bases are indicated by color: green, A; blue, C; brown, G; red, T.
Extended Data Fig. 4 | Postprocessing of putative somatic MEIs. a, Procedure for manual curation of putative somatic MEIs. To further remove false positive MEIs, especially for Alu insertions, we implemented manual inspections for each putative insertion. We first check the neighboring regions in both the UCSC and IGV browsers and remove calls that are from regions of potential mapping errors or CNVs. We also remove calls that are found in datasets of other donors. We then apply a novel visualization tool, RetroVis, to quickly screen out calls with questionable supporting read positions. We further inspect the read sequences to check for unwarranted transduction and similarity between different supporting reads. Finally, we design nested PCR and ddPCR to validate the insertions and quantify their respective levels of mosaicism using DNA from the same tissue. In a RetroVis plot, black lines represent human genome location (top) and the inferred segment of the inserted mobile element (for example, L1) (bottom). A paired-end supporting read is represented by a blue arrow and a red (+ strand insertion) or purple (-strand insertion) arrow connected by a dashed line. A split-read supporting read (spanning an insertion junction) is plotted as a blue arrow (reference segment) connected to an empty rectangle (mobile element segment), with a red or purple arrow below. The positions of the blue segments and red/purple segments reflect the insertion coordinates in the human reference genome and mobile element consensus. b–j, Examples of likely false positive insertions examined by manual curation. Blue, flanking sequence; red, mobile element sequence (+ strand insertion). b, Merging different MEIs into one. c, PCR duplicates. d, All ME ends are mapped to identical coordinates at the 3' end of the L1HS sequence. e, All anchor ends are mapped to identical coordinates in flanking sequences. f, Lacking target site duplication. g, A truncated 3' end indicates a false insertion or an endonuclease-independent retrotransposition. h, Two supporting reads mapping to the same ME location but having a low sequence similarity. i, When the split-read supporting read is mapped partially to the ME consensus (red, locus 2) and fully to another reference genome element (green and red, locus 1), the additional sequence (green) is transduced to the new location. Transduction in Alu insertions, or 5' transduction in 5'-truncated L1 insertions, indicates a false insertion. j, The supporting reads suggest that the ME is inserted in the + strand, yet the 3' end is closer to the upstream flank and the 5' end is closer to the downstream flank. This conflict indicates a false insertion or a 5' inversion in L1 retrotransposition.
Extended Data Fig. 5 | See next page for caption.
**Extended Data Fig. 5 | Summary of the validation experiments.**

**a.** We used droplet digital PCR (ddPCR) to confirm presence of detected somatic L1s in the DNA from combined cells and to measure the tissue allele frequency, and nested PCR to sequence the junctions (1st nested PCR is the reaction containing both ends of the insertion, and the 2nd nested PCR then uses the product of the 1st as template and targets upstream or downstream junctions). **b.** We applied nested PCR to amplify the 5' and 3' junctions for L1-1 and L1-2 with overlapping primers, and then used overlap extension PCR (OE-PCR) to obtain the full sequence of L1-1 and L1-2. Control DNA was amplified on DNA without the L1 insertion (NA12878) using primer iii and primer vi. The amplified DNA (L1 or control) was cloned to a constitutively spliced intron in an enhanced green fluorescence protein (EGFP) reporter, pGint. **c.** An example of biased PCR amplification favoring pre-integration (insertion-) site blocks the amplification of the post-integration (insertion+) site even at relatively high tissue allele frequencies. We titrated the L1-1 template from GL1-1 plasmid in NA12878 genomic DNA at allele frequencies of 92.4%, 64.6%, 20.7%, 3.59% and 0.53%, and then tested PCR amplification with external primers using PhusionTaq or DreamTaq polymerases, and 30 or 60 PCR cycles (n=1 replicate for each PCR cycle). **d.** We designed a droplet-based full length PCR to reduce bias and amplify the post-integration site. We prepared 8 droplet PCR reactions from the genomic DNA of brain or controls: 7 reactions were combined for gel electrophoresis and the last reaction was tested for the probe fluorescence (for example, again ddPCR). NA12878 genomic DNA was used negative control and the known L1-1 or L1-2 templates was tested as positive controls. **e.** The placement of primers (P1+P2) and probe used in the droplet-based full length PCR for L1-1 and L1-2. Primer P3+P2 and P3+P4 were used for in a second PCR to re-amplify the full length insertion of L1-1 and L1-2, respectively.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Experimental validation of L1-1. a, We used droplet digital PCR (ddPCR) to measure the frequency, nested PCR to sequence the junctions, cloning with overlap extension PCR (OE-PCR) to obtain the full length insertion sequence, and droplet-based full length PCR followed by gel electrophoresis or fluorescence read-out to amplify the post-integration site (see Extended Data Fig. 5d). TSD, target site duplication; up, upstream junction; dn, downstream junction. b, DdPCR detected a clear signal for L1-1 in the genomic DNA from right hemisphere superior temporal gyrus, in both neurons (n=8 replicates) and glia (n=8 replicates), but not in the fibroblast (n=8 replicates). Green, droplets containing only RPP30 (internal control); Blue, droplets containing only the L1 junction template; Orange, droplets containing both L1 and RPP30 templates; Black, droplets containing neither L1 nor RPP30 templates. We used NA12878 DNA as a negative control and synthesized DNA with the target L1 junction as a positive control. c, The full sequence of L1-1 based on OE-PCR. Black, flanking sequence; red, inserted L1 sequence; purple, target site duplication; cyan, L1Hs specific alleles; brown, mismatch to the L1Hs consensus. d, Nested PCR results showed L1-1 upstream and downstream junctions amplified specifically in the genomic DNA of right STG (RSTG) but not in NA12878. This experiment was repeated for 4 times and always showed the same results. Yellow arrow, product of pre-integration site in the 1st nested PCR (934 bp); yellow rectangle, gel extraction from the 1st PCR to serve as template in 2nd PCRs; red arrow: upstream junction in 2nd nested PCR (336 bp); blue arrow, downstream junction in 2nd nested PCR (594 bp); NA12878, negative control. e and f, The gel electrophoresis from three independent replicate experiment of the droplet-based full length PCR, confirming the amplification of the L1-1 post-integration site in glia from two brain anatomical regions: LOP—left hemisphere occipital cortex, proximal to STG and LSTG2—a second sample from left hemisphere superior temporal gyrus. NA12878, negative control; L1-1, positive control with known L1-1 junction from plasmid GL1-1. e, Replicate experiment 1. f, Replicate experiment 2 and 3. g, Fluorescence readout of the droplet-based full length PCR was quantified based on a standard curve where L1-1 template (from plasmid GL1-1) is mixed with NA12878 at 4 different allele frequencies: 10.83%, 19.54%, 24.27% and 32.69%. The ratio of positive droplets is positively correlated with the L1-1 template frequency (Pearson’s r=0.99). The blue line marks the linear trend and the surrounding gray area marks the 95% confidence intervals. h, Fluorescence readout (n=2 anatomical regions) of the droplet-based full length PCR confirms the presence of L1-1 in the tested glial cells but shows no signal in the fibroblasts. The results are displayed in 2 dimensions for clearer illustration, with no internal control used for the signal on the X-axis. The ratio of L1-1 positive droplets (blue) over the total number of droplets is indicated in each ddPCR experiment.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Experimental validation of L1-2. a, We used droplet digital PCR (ddPCR) to measure the frequency, nested PCR to sequence the junctions, cloning with overlap extension PCR (OE-PCR) to obtain the full length insertion sequence, droplet-based full length PCR followed by gel electrophoresis or fluorescence ddPCR to amplify the post-integration site, and ddPCR using a Taqman probe crossing its 5’-junction (see Extended Data Fig. 5d). TSD, target site duplication; up, upstream junction; dn, downstream junction. b, DdPCR detected a clear signal for L1-2 in the genomic DNA from right hemisphere superior temporal gyrus, in both neurons (n=10 replicates) and glia (n=10 replicates), but not in the fibroblast (n=10 replicates). Green, droplets containing only RPP30 (internal control); Blue, droplets containing only the L1 junction template; Orange, droplets containing both L1 and RPP30 templates; Black, droplets containing neither L1 nor RPP30 templates. We used NA12878 DNA as a negative control and synthesized DNA with the target L1 junction as a positive control. c, The full sequence of L1-2 based on OE-PCR. Black, flanking sequence; red, inserted L1 sequence; purple, target site duplication; cyan, L1Hs specific alleles; brown, mismatch to the L1Hs consensus. d, Nested PCR results showed L1-2 upstream and downstream junctions amplified specifically in the genomic DNA of right STG (RSTG) but not in NA12878. This experiment was repeated for 4 times and always showed the same results. Notably, we used two different sets of primers in the first PCR for the upstream and downstream junctions. Yellow arrow, product of pre-integration site in the 1st nested PCR (L1-2 up, 266 bp; L1-2 dn, 561 bp); yellow rectangle, gel extraction from the 1st PCR to serve as template in 2nd PCRs; red arrow: upstream junction in 2nd nested PCR (263 bp); blue arrow, downstream junction in 2nd nested PCR (215 bp); NA12878, negative control. e, Gel electrophoresis of the droplet-based full length PCR confirmed the amplification of the L1-2 post-integration site in neurons from the right hemisphere occipital cortex, distal to STG (ROD). NA12878, negative control; L1-2, positive control with known L1-2 junction from L1-2 OE-PCR (see Extended Data Fig. 5b). The droplet-based full length PCR experiment was repeated and showed similar results. f, Fluorescence readout (n=1 replicate) of the droplet-based full length PCR confirms the presence of L1-2 in neurons from ROD but shows no signal in the fibroblasts. The results are displayed in 2 dimensions for clearer illustration, with no internal control used for the signal on the X-axis. The ratio of L1-2 positive droplets (blue) over the total number of droplets is indicated in each ddPCR experiment. The quantification of the L1-2 frequency is based on a standard curve where L1-2 template (from L1-2 OE-PCR) is mixed with NA12878 at allele frequencies of 7.25% and 13.51%.
Extended Data Fig. 8 | Spatial distribution and poly(A) length of L1-1 and L1-2. a, Anatomical brain regions studied in donor 12004: 1 and 1', superior temporal gyrus (BA22, both sides); 2, prefrontal cortex distal (BA9, both sides); 3, prefrontal cortex proximal (BA46, both sides); 4, motor cortex distal (BA4, both sides); 5, motor cortex proximal (BA6, both sides); 6, parietal cortex distal (BA7, both sides); 7, parietal cortex proximal (BA39, both sides); 8, occipital cortex distal (BA19, both sides); 9, occipital cortex proximal (BA19, both sides); 10, putamen (both sides); 11, cerebellum (both sides). The tissue for deep whole genome sequencing is from right superior temporal gyrus (1'). The tissues that were dissected from both hemispheres were bilaterally symmetrical. The metric unit on the ruler is the centimeter. b, The levels of mosaicism in neurons are highly correlated with levels in glia. Red, L1-1; green, L1-2. c, Poly(A) lengths of L1-1 and L1-2 were estimated as the lengths supported by the highest numbers of GL1-1 and GL1-2 clones (see Supplementary 8b). The variation among clones was likely the result of PCR stutter around low-complexity templates67. d, Poly-A length distribution in 22 previously reported de novo and disease-causing L1 retrotranspositions. The poly-A lengths of L1-1 and L1-2 are at 18.2% and 13.6% percentiles, respectively, of this distribution.
Extended Data Fig. 9 | The genomic locus with L1-1 insertion. L1-1 is inserted in a 2.6 kb promoter flanking region (ENSR00000032826) that is hypothesized to regulate the expression of nearby genes. The chromatin states are shown for a subset of human cell lines: light gray, heterochromatin; light green, weakly transcribed; yellow, weak/poised enhancer; orange, strong enhancer; light red, weak promoter; bright red, strong promoter. L1-1 is inserted in a linkage disequilibrium (LD) block, based on the common SNPs that are highly correlated ($R^2 > 0.6$, green line) with the closest common SNP to L1-1, rs1890185. This LD block is highlighted in red, and contains 72 lead SNPs associated with 10 diseases or disorders and 28 measurements or other traits, including 13 risk SNPs from 11 schizophrenia studies (triangle). We categorized all traits under 11 terms based on the Experimental Factor Ontology. The significantly associated SNPs, indexed from number 1 to 72, are documented in details in Supplementary Table 6.
Extended Data Fig. 10 | Fluorescence quantification in the reporter assay.  

**a-b.** Original photos of the representative images in Fig. 6d,e.  
**c.** Raw fluorescence intensities (green and red) used in the statistical analysis in Fig. 6f,g were in the range of 0–3035 for green fluorescence and 0–3613 for red, with no saturated pixels (>4000). Each cell is represented by the average pixel intensity (dot) and the maximum and minimum pixel intensities (bar). Red, Gcont-1; Cyan, GL1-1; Green, Gcont-2; Purple, GL1-2.  
**d.** Measurement of the green fluorescence, red fluorescence and brightfield of three cells. C1, live cell; C2, dead cell, C3, dead cell. Each image is a representative of the green and red fluorescence images in well 1 to well 5 for any reporters (total=60).  
**e.** Representative images from each the GFP fluorescence of the control and L1-1 reporters in the single transfection experiment (2 wells and 3 images per well, see Fig. 6c). The maximum signal intensities are adjusted from 4095 to 1000 in (d) and (e) to illustrate the cells with weak fluorescence.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about: availability of computer code

Data collection
The droplet digital PCR data were collected with QuantaSoft Analysis Pro Software (v1.0, BioRad). The reporter fluorescence intensities were quantified with Leica Application Suite 300 (build 8134).

Data analysis
1. Sequencing read alignment:
   Burrows-Wheeler Aligner (BWA v0.7.12), samtools (v1.9), picard (v1.92) and bedtools (v2.27.1)
2. Candidate Supporting reads for mobile element insertions
   RetroSeq (v1.43)
3. Machine learning:
   R (v3.5.0), including glmnet package (v2.0-16), randomForest package (v4.6-14), e1071 package (v1.6-8) and ggplot (v3.2.0)
4. RetroSom (v1), RetroVis (v1) and plotting the main figures:
   The code is available in the supplementary software file and at https://github.com/XiaoweiZhuJJ/RetroSom

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The whole genome sequencing data of the six donors (Fig. 1a, b) have been deposited in Sequence Read Archive under BioProject ID: PRJNA541510 (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA541510). These data are publicly available.

The source data for the genome-mixing experiment (Fig. 2c) are deposited in the NIMH Data Archive (https://nda.nih.gov/) under Collection 2458, Experiment 1072. The data are not publicly available due to them containing information that could compromise research participant consent, but will be available from the corresponding author upon reasonable request.

The Microscope image collection for the reporter assay are available in Figshare collection 5182676: https://doi.org/10.6084/m9.figshare.c.5182676.v1. These images are publicly available.

Source data are provided for Fig. 1,2,3,4, and 6. Original gel images are provided for the Extended Data Fig. 3, 5, 6 and 7.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Sample size**
We applied deep whole genome sequencing for neurons, glia and a non-brain tissue in six donors. The samples in this study (e.g., 6 human brains) were not utilized to determine the correlation between somatic retrotransposition and Schizophrenia, and the sample size is similar to those reported in previous studies in discovering the rates of somatic retrotranspositions.

**Data exclusions**
To call somatic mobile element insertions, we excluded supporting reads of poor quality based on pre-established criteria, including (1) genomic regions of highly repetitive sequences, including centromeric repeats, telomeric repeats, large segmental duplications, reference genome gaps, or within 100bp of a reference MEI of the same type and strand; (2) supporting reads with low sequencing complexity (SEG score < 1); or (3) outlier sequencing depth within 500bp upstream and downstream to the insertion (>3 standard deviations away from the mean). The sequencing depth for sex chromosomes was evaluated separately. The masked reference sequence was 23.6% for L1 insertions in the positive strand, 23.7% for L1 insertions in the negative strand, 21.0% for Alu insertions in the positive strand, and 21.1% for Alu insertions in the negative strand.

**Replication**
(1) The initial finding of L1-1 and L1-2 were validated in droplet digital PCRs with more than four technical replicates, and the results confirmed their presence in ~1% of the neurons and glia, but not in fibroblast or negative control (NA12878 genomic DNA) (Figure 3b, 3e).
(2) The presence of somatic L1-1 and L1-2 were further validated with (i) nested PCR, (ii) overlap-extension PCR, and (iii) droplet-base full length PCR (Extended Fig. 5-7).
(3) The ‘Gint’ reporter assay was tested with 2 separate sets of replicate experiments, with and without co-transfection of the internal control (‘Rint’) (Figure 6b, 6c). The changes of fluorescence were consistent in both experiments (Figure 6f, 6h).
We have added specific numbers of replicate experiments in the related figure legends.

**Randomization**
To ensure the consistency and complexity of the sequencing libraries, we prepared 6 separate sequencing library for each tissue type of each donor, sequenced each library for average depth >30x for a total sequencing depth of 200x.
In the reporter assay (Figure 6), the labels of the L1 and control reporters were randomized during the transfection experiment.

**Blinding**
The status of schizophrenia/control of the two pairs of donors (10011, 11003, 11004 and 12004) is hidden from the entire team at Stanford until the somatic mobile element insertions were called and validated.
In the reporter assays (with and without the internal Rint control), the order of each transporter assay was shuffled and kept hidden until the fluorescence was quantitated by a different experimenter.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
Antibodies

Antibodies used
- Anti-NeuN-PE (Milli-Mark, FCMA8317PE, clone A60, lot #3065043 and #3153277); anti-CD45 (BD 550539, clone 30-F11), anti-HepaCAM (R&D MAB4108, Clone # 419305) anti-Thy1 (BD 550402, clone SE10)

Validation
1) FCMAB317PE antibody was used in previous studies for sorting human neuron and non-neuron nuclei (e.g., PMID: 30263963 and 30973874). We used the same antibodies (anti-CD45, anti-HepaCAM and anti-Thy1) in immunopanning as described in a previous study on separating cell types in human fetal brain (PMID: 26687838).
2) The manufacture of FCMAB317PE evaluated the quality by flow cytometry using U251 cells. The sorted neuron nuclei in this study is also consistent with several neuron hallmarks, including 1) size is larger than non-neuron nuclei and 2) the ratio of neuron and non-neuron counts is the largest in cerebellum, ~1:2 in cortical regions and much lower in subcortical regions (Supplementary Fig. 1).
3) The 30-F11 clone has been reported to react with all isoforms and both alloantigens of CD45, which is found on hematopoietic stem cells and all cells of hematopoietic origin, except erythrocytes. CD45 is a transmembrane glycoprotein which is expressed at high levels on the cell surface, and its presence distinguishes leukocytes from non-hematopoietic cells. CD45 is a member of the Protein Tyrosine Phosphatase (PTP) family, where the intracellular carboxy-terminal region contains two PTP catalytic domains, and the extracellular region is highly variable due to alternative splicing of exons 4, 5, and 6 (designated as A, B, and C, respectively). CD45 isoforms play complex roles in T-cell and B-cell antigen receptor signal transduction and the CD45 isoforms detected in the mouse are cell type-, maturation-, and activation state-specific.
4) Anti-HepaCAM (R&D MAB4108) detects human HepaCAM in direct ELISAs and Western blots.
5) The 5E10 monoclonal antibody specifically binds to human CD90 which is also known as Thy-1. CD90 is a 25-35 kDa glycophasphatidylinositol-anchored membrane glycoprotein of the Ig superfamily that is expressed on 1-4% of human fetal liver cells, cord blood cells, and bone marrow cells. The anti-CD90 antibody binds to a subset of immature CD34+ cells and a distinct subset of mature CD34- cells that are CD3+CD4+. The CD90+CD34+ population is highly enriched for cells capable of long-term culture. The anti-CD90 antibody is useful for enriching high proliferative potential colony-forming cells (HIPP-CFC) that are primitive progenitor cells.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
- We used four fibroblast cell lines collected via dermal biopsy from the upper arm of donor "10011", "11003", "11004" and "12004"

Authentication
- We tested each cell line with >200x whole genome sequencing, and confirmed that they carried the same germline mobile element insertions as the corresponding brain cell, hence we confirm the labeling of the tissues is correct.

Mycoplasma contamination
- The cell lines were not tested for the Mycoplasma contaminations. The fibroblast cell lines were used as controls for the brain tissues for excluding germline mobile element insertions. In addition, we studied human specific mobile elements (AluY and L1Hs), which are not present in Mycoplasma bacteria.

Commonly misidentified lines
- No commonly misidentified cell lines were used.

Human research participants

Policy information about studies involving human research participants

Population characteristics
- The gender, age and other relevant information of all donors are summarized in supplementary table 1. The 2 pairs of SZ and controls were initially recruited for a separate study which was cited in the sample collection section in the methods (Pubmed ID: 24912493). The SZ donors and controls were matched as closely as possible for age, brain pH, postmortem delay to autopsy and RNA integrity number. They are all of the Caucasian ethnicity, as confirmed by a principal component analysis as shown in Supplementary Fig. 5a.

Recruitment
- Specifically for the SZ-control donor pairs, volunteers with a DSM-IV diagnosis of SZ and sixteen matched normal controls (NC) with no history of major psychiatric disease were recruited from the Dallas metropolitan area. Inclusion criteria for all subjects were: English language fluency, competence to give informed consent, age between 18 and 60 years, and good medical health. Exclusion criteria for all volunteers consisted of: pregnancy, any organic brain disease, significant medical illness, history of severe head trauma and current use (within one month) or extensive history of illicit drug use. The fifth
donor, A1S, was recruited based on ethnicity (Caucasian) and no previous history of mental disorders.

Ethics oversight

For the SZ-control pairs, informed consent was obtained for all participants in accordance with procedures approved by the University of Texas Southwestern Medical Center Institutional Review Board. For donor A1S and F1, we obtained postmortem brain tissue and heart tissue after review of the proposed procedures by the Stanford University Institutional Review Board which determined that they did not constitute human subjects research (exempt because research was not performed on living subjects).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

For the initial whole genome sequencing screening of the adult donors, we sampled 0.5-1 cm³ cortical tissues from the superior temporal gyrus (STG). The neuronal and glial nuclei were extracted from the postmortem brains using methods modified from a published protocol2. Briefly, the brain tissues were dissected on a cold plate (TECA™ LHP-1200CAS) into ~200mg segments. For each segment, we homogenized the tissue in 3.6ml lysis buffer (0.32M sucrose, 5mM calcium chloride, 3mM magnesium acetate, 0.1mM EDTA, 1mM DTT, 0.1% TritonX-100, and 10mM Tris PH 8.0). We then added 6.5ml sucrose buffer (1.8M sucrose, 3mM magnesium acetate, 1mM DTT and 10mM Tris PH 8.0) to the bottom of the tissue lysate, and centrifugated at 100,000g for 2 hours at 4 °C (Sorvall™ ultracentrifuge WX-80). The nuclei in the pellet were collected by incubation in 500 μl ice-cold PBS for 10 min, gentle resuspension, and filtration through a 40 μm strainer. We stained the nuclei with an anti-NeuN-PE antibody (Milli-Mark FCMAB317PE, 1:100), 1mg/ml DAPI (1:1000), and 10%BSA (1:50) for 45 min at 4 °C. The labeled nuclei were evaluated under a fluorescent microscope (EVOS FL), and the yield was quantitated with a hemocytometer.

Instrument BD Aira II sorter (https://facs.stanford.edu/instruments/falstaff)

Software The data were analyzed with FlowJo cell analysis software (v10.0.7.r2).

Cell population abundance

A typical yield from 200mg of brain tissue is 1-2 million nuclei, NeuN+ and NeuN- combined. The ratio between the NeuN+ and NeuN- fraction varies depending on the anatomical region, e.g., 1.6 in superior temporal gyrus, 12.6 in cerebellum, and 0.24 in putamen. The purity of the sorted nuclei (quantitated by reanalyzing the sorted fractions) was >99.95% in both fractions.

Gating strategy

We first drew gates in forward scatter (FSC-A and FSC-W), side scatter (SSC-A and SSC-W), and DAPI channels to select for singlet nuclei. The NeuN+ and NeuN- nuclei were then separately collected with gates in the PE and FSC-A channels: NeuN+ nuclei are from neurons and are larger in size and carry stronger PE signals, while NeuN- nuclei are from non-neurons (glial cells) and are smaller. The gating strategy is provided in Supplementary Fig. 1.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.