

Supplementary Results

RC-seq performance and reproducibility

Of the 177.4 million RC-seq reads produced from seven brain libraries, 1.34 million aligned with high confidence to targeted capture array retrotransposons and 98.2% of the remainder corresponded to other closely related transposed elements (Supplementary Table S3). This represented a 20-fold enrichment for target sequences.

Qualitative and quantitative performance of RC-seq was assayed using the two replicates of donor A caudate nucleus. Overall, 772/780 (99.0%) of retrotransposons targeted by the capture array and detected by a diagnostic read in Replicate 1 were detected in Replicate 2 (Supplementary Table S10). Quantitative reproducibility was also very high overall, fitting a linear trend line of $y=1.002x$, $R=0.95$ (Supplementary Fig. S4). Using Replicate 1 as an example, 89.1% of all targeted elements were detected, including 94.2% of L1s (Supplementary Table S11). Many of the remainder were likely to be true-negatives corresponding to polymorphisms¹ or were refractory to detection due to the presence of other repetitive elements.

Again using caudate nucleus from donor A as a representative example, 71 diagnostic reads aligned on average to each L1 and 60.7% of L1s were identified at both termini, with corresponding statistics of 18 reads and 68.8% for *Alu* (Supplementary Table S2, Supplementary Table S11). Capture and sequencing bias were limited; the composition of the capture array (Supplementary Fig. S5a) and the proportion of reads aligned to each retrotransposon family (Supplementary Fig. S5b) were similar except for a moderate increase in LTRs at the expense of *Alu*. Altogether, these statistics indicated that RC-seq reliably captured targeted elements present in the human reference genome².

Comparison of L1 and Alu somatic retrotransposition rate

The human genome contains an estimated 100 transposition-competent L1s³ and ~3,000 active *Alu* elements^{4,5}. We found 7,743 somatic L1 insertions and 13,692 somatic *Alu* insertions. Therefore, the per element somatic retrotransposition rate observed here was approximately 80:1 and 4:1 for L1 and *Alu*, respectively. It is important to note that these ratios can only be used to compare L1 and *Alu*. Each somatic insertion is present in a minority of cells, meaning that, for example, the somatic:germ line L1 insertion ratio in individual cells is much lower than 80:1.

L1 and Alu retrotransposition produce somatic mosaicism

Retrotransposons restructure the genome through several mechanisms⁶. Previous observations⁷⁻⁹ indicate that L1 retrotransposition, rather than an alternative process such as non-homologous recombination, is primarily responsible for somatic mosaicism in the brain. Several outcomes in the current study strongly support this position. We observed extensive L1 CNV between brain regions that was unlikely to arise from recombination⁷, traced 80.2% (Supplementary Fig. S3a) of somatic L1 insertions to the two L1 subfamilies most likely to mobilize in germ cells⁵ (L1-Ta and pre-Ta) and found that somatic L1 insertions disproportionately affected euchromatic regions susceptible to nicking by the L1 endonuclease¹⁰. L1 retrotransposition via target-primed reverse transcription (TPRT) typically involves a donor element with a polyA tail and produces target site duplications (TSDs) upon insertion⁶. One somatic L1-Ta insertion was sequenced at both its 5' and 3' termini by RC-seq, indicating the presence of a TSD at least 73bp in length and a suitable L1 endonuclease nick site. To infer the overall proportion of somatic L1 insertions incorporating a polyA tail, we performed long paired-end 150mer sequencing on donor B samples to assemble the 3' junction sequence of 30 somatic L1 insertions. Twenty-seven (90.0%) presented a clear polyA tail (median length: A_{23}).

With respect to *Alu*, 83.0% of somatic insertions involved the AluY subfamily (Supplementary Fig. S3b) that is most likely to cause dimorphic insertions among humans⁵. 37.9% corresponded to AluYa5 or

AluYb8. All of the remaining *Alu* insertions reported here were AluS, a subfamily known to be active in cultured cells¹¹. Using the same strategy applied to L1, 245/334 (73.4%) of all somatic *Alu* insertions found by RC-seq on donor B samples contained a polyA tail (median length: A₁₀). Two somatic *Alu* insertions were detected at both their 5' and 3' termini, with one of these events (an AluY) producing a TSD of 44bp and flanked by an L1 endonuclease nick site. Given that efficient *Alu* mobilization by the L1 machinery is correlated with polyA tail integrity¹²⁻¹³ and often coincides with TSD production by TPRT, these results provided further evidence that retrotransposition drives *Alu* mobility in the brain.

Somatic SVA mobilization

In comparison to L1 and *Alu*, there were fewer non-reference SVA insertions detected by RC-seq. SVAs accounted for 3.9% of capture array probes (Supplementary Fig. S5a), 2.6% of reads aligned to targeted reference genome elements (Supplementary Fig. S5b) and 5.6% of non-reference insertions (Fig. 3a). SVA, like *Alu*, is thought to be mobilized in germ cells by the L1 machinery¹⁴. To assess whether this had occurred in the brain, we extended the nested PCR strategy used for L1 and *Alu* to four SVA insertions (Supplementary Table S7). Two SVA_F insertions were confirmed as somatic and one SVA_E insertion as germ line. Both somatic insertions arose from near full-length elements. This suggested that a handful of SVAs may have mobilized in the brain samples studied here.

Non-reference LTR insertions

Although LTRs accounted for 22.2% of capture array probes (Supplementary Fig. S5a) and 38.4% of reads aligned to targeted reference genome elements (Supplementary Fig. S5b) only 1.3% of novel insertions detected by RC-seq were LTRs (Fig. 3a). This massive depletion in the number of novel LTR insertions contrasted strongly with results for L1, *Alu* and SVA and suggested that the selected LTRs had functioned properly as negative controls. We initially considered that most if not all non-reference LTR insertions were false positives. To confirm this, we tested four examples by nested PCR. Only one of the four validated, as a somatic insertion (Supplementary Table S7). It was plausible that an LTR containing a polyA tail could be mobilized by L1. However, the one confirmed somatic LTR insertion did not present an obvious polyA sequence. We concluded that the handful of non-reference LTR insertions detected by RC-seq were mainly false positives, germ line events or somatic rearrangements not involving retrotransposition.

Supplementary Discussion

RC-seq can detect rare somatic retrotransposition events

Several high-throughput methods have recently been used to survey germ line retrotransposition events in humans¹⁵⁻¹⁸. One study identified somatic L1 insertions in lung cancer¹⁸ which, due to clonal replication during tumorigenesis, could affect numerous cells in a given tumor. By contrast, RC-seq successfully detected rare somatic events in subsets of cells in apparently normal tissue. A sequence capture-based enrichment strategy circumvented PCR amplification of active retrotransposons and used fewer cycles of PCR overall for library preparation than other methods¹⁵⁻¹⁸.

Somatic mutations in minor cell subpopulations may be overlooked after numerous rounds of PCR as a competitive reaction could favor comparatively abundant germ line insertions. This distinction is also the biggest drawback of using sequence capture against repetitive elements; substantial background is generated by cross-hybridization and must be overcome by deep sequencing. Our termini-focused sequence capture approach also likely underestimated the frequency of 5' truncated L1s more than 1kb in length. Future gains in efficiency may be achieved through improved probe selection, liquid-phase sequence capture and ongoing reductions in sequencing costs.

Supplementary Methods

Sample details and DNA extraction

Tissues were provided by the Netherlands Brain Bank (Amsterdam, The Netherlands) for three donors

(donor A, male, 91 years; donor B, male, 87 years; donor C, female, 97 years). Neuropathological examination revealed negligible neurofibrillary tangles, senile plaques or Lewy bodies. Dissection of the caudate nucleus, putamen, hippocampus, middle frontal gyrus and middle temporal gyrus was performed from human brain sections. Genomic DNA was extracted and purified from each tissue using the Gentra Puregene Tissue Kit (Qiagen, Cat# 158667) according to instructions provided by the manufacturer. DNA was visualized by gel electrophoresis, presented a minimum 260/280 absorbance ratio of 1.7 and quantified by nanodrop. Pooled human male genomic DNA was purchased from Promega (Cat# G1471) and extracted without cell culture from leukocytes corresponding to multiple anonymous whole blood donations.

TaqMan qPCR

Quantitative PCR experiments were performed on a Roche LightCycler 480 with minor modifications to the approach of Coufal et al.⁷. In each PCR, 1ng of genomic DNA was combined with gene-specific primers (Invitrogen), a TaqMan-MGB probe (Applied Biosystems) and 12.5μl LightCycler 480 Probes Master (Roche NimbleGen, Cat# 04707494001) in a total volume of 25μl. For the L1 assay, primer concentrations were 0.9μM and TaqMan probe concentration 0.3μM. For the α-satellite (SATA) assay, primer and TaqMan probe concentrations were 0.4μM. PCR conditions were as follows: 95°C for 5 minutes followed by 45 cycles of 95°C for ten seconds and 59°C for one minute. Standard curves of genomic DNA ranging from 200ng to 0.2ng were performed to verify that the 1ng dilution was within the linear range of the reaction. Quantification included five technical replicates. For each assay, the ratio of L1 ORF2 to SATA was normalized to the ratio obtained for caudate nucleus. Primers and TaqMan probes were identical to those designed by Coufal et al.⁷, with the “ORF2 #1” combination used to assay L1. For each individual, ORF2/SATA ratios were compared across five brain regions with a repeated measures one-way ANOVA followed by pairwise LSD post hoc testing and Bonferroni correction.

Retrotransposon capture array design

Eight hundred and seventy-five full-length L1, *Alu* and SVA retrotransposons, as well as putatively immobile¹⁹ ERV1 and ERVK LTRs were selected from the human reference genome (hg19) using RepeatMasker coordinates²⁰ and stringent identity cutoffs (Supplementary Table S1). A list of these elements is provided in Supplementary Table S2. For L1, SVA and LTR elements, oligonucleotide probes were tiled across the two 300bp regions immediately within their 5' and 3' termini, without filtering for repetitiveness. *Alus* were tiled in full due to their short length (~300bp). In most cases, therefore, eight probes were generated per L1, SVA and LTR and four probes per *Alu*, with a total of 4,885 probes (Supplementary Table S2). These features were replicated to fill a custom NimbleGen Sequence Capture 2.1M Array (Roche NimbleGen).

Capture library preparation and pre hybridization amplification

Paired-end libraries were constructed using an Illumina kit (Illumina Inc., Cat# PE-102-1001) and the following modifications: the prescribed agarose gel excision was performed at 300-350bp to produce libraries with an approximate fragment size of ~375bp. DNA was purified from agarose using a Qiagen QIAquick column (Cat# 28104) and eluted in 30μl of water. This template was amplified by ligation mediated PCR (LM-PCR) using Illumina sequencing adapters in a reaction containing 50μl Phusion High Fidelity PCR Master Mix (New England BioLabs, Cat# F-531L), 2μM Illumina PE primers 1.0 and 2.0, 30μl DNA and water up to 100μl. PCR conditions were as follows: 98°C for 30 seconds, followed by 8 cycles of 98°C for 10 seconds, 65°C for 30 seconds and 72°C for 30 seconds, and final extension at 72°C for 5 minutes. Amplified material was then cleaned with a Qiagen QIAquick column following manufacturer's instructions, except DNA was eluted in 50μl water. DNA was quantified using a NanoDrop-1000 and insert size confirmed with an Agilent Bioanalyzer 2100 and DNA1000 chip (Cat# 5067-1504). The Illumina PE primer sequences were:

1.0: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'

2.0: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC*T-3'

(*denotes a phosphorothioate bond)

Capture library and capture array processing

Libraries were processed prior to array hybridization by adding the following components to a 1.5ml tube: 320ng of library material, 0.64µl of 100µM Illumina PE primers 1.0 and 2.0 and 64µl of Roche NimbleGen proprietary capture enhancing (PCE) compound (available upon request). Samples were dried down and then re-suspended by adding 15.4µl of water and placing tubes on a heating block at 70°C for 10 minutes, followed by vortexing for 30 seconds and centrifugation. To each sample tube, 25.6µl NimbleGen SC Hybridization Buffer and 10.24µl NimbleGen Hybridization component A (Cat# 05340721001) were added, followed by vortexing for 30 seconds, centrifugation and placement on a heating block at 95°C for 10 minutes. Samples were again mixed for 10 seconds, spun down, and placed in a Roche NimbleGen Hybridization System at 42°C until ready for hybridization. The capture array consisted of three identical sub arrays of 720K features. To each slide, a NimbleGen HX3 mixer was affixed according to the manufacturer's instructions and 16µl of the hybridization mixture (sample library, PCE, Illumina primers, SC Hybridization Buffer and SC Component A) was pipetted into each of the three sub-array fields. Loading and vent holes were covered with port seals and each sample was hybridized for 72 hours at 42°C on Hybridization Station setting "B". Slide washing and sample library elution were performed as previously described²¹.

Post hybridization ligation mediated PCR

Post hybridization LM-PCR amplification consisted of 2 reactions for each sample using the same enzyme and primer concentrations as the pre capture amplification and different Illumina PE primers (1.0: 5'-AATGATACGGCGACCACCGAGA-3' and 2.0: 5'-CAAGCAGAAGACGGCATACGAG-3'). Post hybridization amplification consisted of 16 cycles of PCR with identical cycling conditions as used in the pre hybridization LM-PCR, except the annealing temperature was lowered to 60°C. Following amplification, samples were purified using a Qiagen QIAquick column. DNA was quantified using a NanoDrop-1000 and again evaluated with an Agilent Bioanalyzer 2100 and a DNA1000 chip.

Capture validation by qPCR

Target enrichment was initially confirmed by qPCR against the *Alu* consensus (forward primer: 5'-TCCTGCCTCAGCCTCCCAAG-3', reverse primer: 5'-GTCAGGAGATCGAGACCATCCC-3'). Pre and post capture samples were diluted to 5ng/µl and 1µl was used in the following reaction: 7.5µl Roche Applied Science LightCycler 480 Sybr Green I Master Mix (Cat# 04887352001), 0.3µl of 2µM forward and reverse primers, and 5.9µl PCR grade water for a total volume of 15µl. The following qPCR cycling conditions were used: 95°C for 15 minutes followed by 40 cycles of 95°C for 10 seconds and 60°C for 1 minute. Products were then validated using melt curve analysis, beginning at 60°C and ending at 95°C. Crossing point (Cp) values were calculated using Roche Applied Science LightCycler 480 Software release 1.5.0 second derivative maximum method (Cat# 04994884001). Post hybridization Cp values were subtracted from pre hybridization Cp values and the resulting value (Δ Cp) was raised to the power of the primer efficiency to calculate fold enrichment. Mean fold enrichment across all samples was 41.7.

Capture library sequencing

Retrotransposon capture library concentration was quantified using a Kapa Library Quantification Kit (Kapa Biosystems, Cat# KK4824) and diluted to 6.6pM for each sample, followed by loading on an Illumina v4.0 paired-end flow cell using the standard Illumina Cluster Station protocol. Sequencing was performed by ARK-Genomics, The Roslin Institute, on an Illumina GAIIx instrument using the standard paired-end sequencing protocol and SCS2.6. Images were analyzed using RTA1.6 to create intensity and .qseq files, followed by conversion to .fastq by GERALD. A total of 164,031,856 2x101mer paired-end reads were initially produced from six brain and blood libraries followed by 38,384,481 2x150mer paired-end reads from two additional brain libraries (Supplementary Table S3).

Paired-end read mapping

Read quality control was performed to minimize error rate whilst maximizing read length. Reads were only included if their initial 33mer presented an average quality score ≥ 10 and fewer than two bases with a quality score < 5 . Sequences were then extended until the average quality score dropped below 10, two bases had a quality score < 5 or a full 101mer (or 150mer) was produced. The resulting mean read length exceeded 80nt in all libraries and 94.6% of read pairs passed QC (Supplementary Table S3). Thirty-three nucleotides was selected as the minimum read length to ensure a high level of specificity with two errors permitted, based on previous calculations²².

Next, reads were aligned to hg19 using SOAP2²³ with only the best alignment(s) recorded, a maximum of two errors per alignment and no limit upon the number of alignments (options $-M\ 4 -r\ 2 -v\ 2$). Alignments were post-processed to identify read pairs with extreme termini separated by less than 500bp, on opposing strands and sequenced towards one another on the genome. Pairs that overlapped a transposed element were then stratified into those that mapped multiple times on the genome (MuMs) and those that mapped to a single location (SiMs). The latter group was further divided into those possessing one uniquely mapped end (“diagnostic” SiMs) and other SiMs (Supplementary Table S3).

Detection of non-reference retrotransposon insertions

Unpaired reads where both ends mapped to the genome were first identified. Given a median insert size of ~250bp (Supplementary Table S3) we also expected a minor but important cohort of individual read ends to overlap retrotransposon insertion breakpoints. Therefore, unmapped reads paired with an aligned read were truncated to a 33mer and re-aligned to hg19 using SOAP2 (options $-M\ 4 -r\ 2 -v\ 1$). If now aligned to the genome but not adjacent to a full-length partner, these truncated reads were retained. All mapped but still unpaired read pairs were compiled and filtered to keep those where one end mapped uniquely to the genome and the other to a RepeatMasker annotated transposed element. These were considered diagnostic for novel retrotransposon insertions.

Rigorous filtering was applied to remove false positives from this list. Firstly, read pairs were annotated by BLAT²⁴ at 95% identity against representative L1-Ta, pre-Ta and L1PA2 sequences³ and AluY, AluS, SVA and LTR consensus sequences obtained from RepBase (www.girinst.org). One end was required to map in full to a consensus sequence or, due to an insertion breakpoint, produce a truncated alignment at the 5' or 3' terminus in the correct orientation. In some cases, reads aligned to more than one retrotransposon subfamily equally well. Ties were resolved as follows: L1: L1-Ta > pre-Ta > L1PA2, Alu: AluYa5 > AluYb8 > AluY (other) > AluS, SVA: SVA_F > SVA_E > SVA_D.

Read pairs were then filtered against hg19 using BLAT. Pairs were removed if their uniquely mapped end aligned at $\geq 90\%$ identity to a locus other than the one found by SOAP2. Reads were also removed if they could be successfully paired in the correct orientation within 1kb of each other based on BLAT alignments at $\geq 80\%$ identity. The purpose of these BLAT alignments was to omit pairs where alignment at lower identity contradicted the initial assessment based on SOAP2. Finally, pairs where both ends mapped to one retrotransposon family (e.g. an *Alu* within another *Alu*) were removed.

Pairs were clustered if their uniquely mapped ends occurred within 100bp on the same strand of the reference genome. Clusters were annotated with respect to their position in protein-coding (RefSeq) genes, the family of the donor retrotransposon and whether a germ line insertion had previously been observed at that locus^{1,16,18}. Only clusters where the donor retrotransposon family and strand in relation to the donor retrotransposon were identical for $\geq 90\%$ of the constituent pairs were retained. Subfamily annotations were assigned based on the aforementioned alignment to consensus sequences following a majority rule. Reads occurring in clusters were counted as diagnostic to novel insertions (Supplementary Table S3). Sense-antisense clusters occurring within 250bp on the genome were recorded as detecting two ends of the same insertion if the donor retrotransposon family was identical and the strands on that donor retrotransposon were opposing in the two clusters. A complete list of annotated novel insertions is provided in Supplementary Table S4.

Identification of polyA tails

RC-seq from donor B was used (paired-end 150mer, median insert size ~200nt) to assemble pairs across 3' junction sequences and identify polyA tails. A minimum overlap of 10nt with >80% was required to join each end of a read pair. Only read pairs corresponding to the forward strand of donor elements were analyzed. A minimum polyA sequence of A₆ was required at >80% identity.

Confirmation of germ line insertions by PCR

Thirty-five novel retrotransposon insertions detected by RC-seq were assayed with PCR using insertion site and retrotransposon specific primers designed with Primer3²⁵. PCR reactions contained 0.1μL GoTaq hot-start DNA polymerase (Promega, Cat# M5005), 1X PCR buffer, 1mM MgCl₂, 0.2mM dNTP, 1μM of each primer and 1ng template DNA in a 25μL tube. The following cycling conditions were used: 94°C for 4 minutes, then 35 cycles of 94°C for 1 minute, 55°C for 1 minute, 72°C for 30 seconds, followed by a single extension step at 72°C for 10 minutes. Optimization in some cases required adjustments to annealing temperature, MgCl₂ concentration or DNA template concentration (see Supplementary Table S6 for details). All products were capillary sequenced using an ABI3730 (GenePool, Edinburgh) to confirm sequence specificity.

Confirmation of somatic insertions by PCR

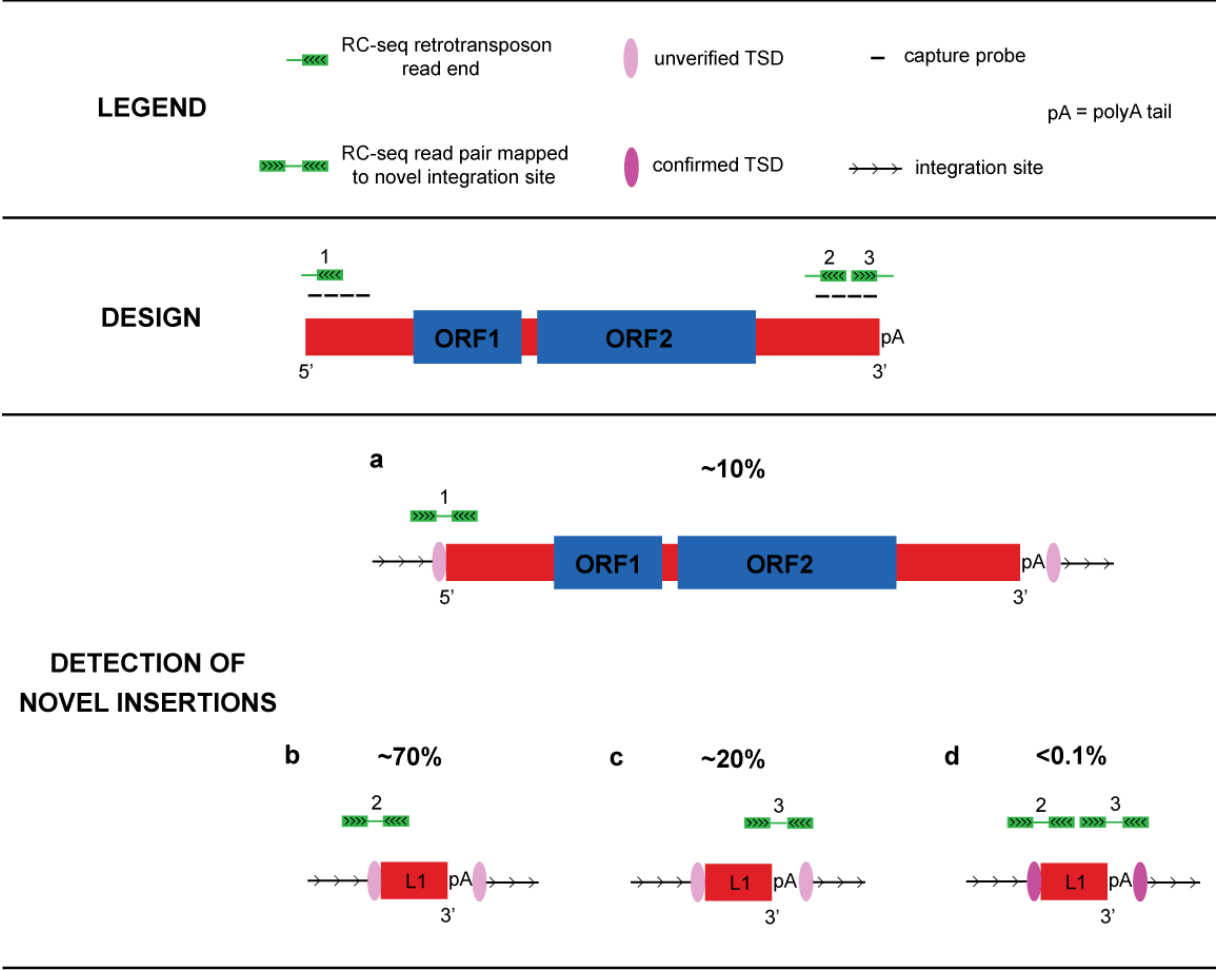
Nested insertion site and retrotransposon specific primers were designed for putative somatic L1, *Alu*, SVA and LTR insertions using Primer3²⁵ following the design laid out in Fig. 4. First round PCR reactions contained 0.5μL AccuPrime hot-start DNA polymerase (Invitrogen, Cat# 12339-016), 1X PCR Buffer II, 1μM of each outer primer and 100ng post-hybridization template DNA in a 25μL tube. The following cycling conditions were used: 94°C for 3 minutes, then 35 cycles of 94°C for 30 seconds, 57°C for 30 seconds, 70°C for 30 seconds, followed by a single extension step at 70°C for 10 minutes. Second round PCR reactions contained 0.5μL MyTaq hot-start DNA polymerase (Bioline, Cat# BIO-21112), 1X PCR buffer, 1μM of each inner primer and 2 μL first round product in a 25μL tube. The following cycling conditions were used: 95°C for 2 minutes, then 25 cycles of 95°C for 15 seconds, 56°C for 15 seconds, 72°C for 10 seconds, followed by a single extension step at 72°C for 10 minutes. Products were capillary sequenced using an ABI3730 (GenePool, Edinburgh) to confirm sequence specificity. All experiments were performed in duplicate with the same products observed for each validated example in both replicates.

Several developments were required beyond the basic approach that succeeded in verifying germ line insertions. In addition to a nested PCR strategy and different polymerases, initial confirmation was only achieved through the use of 100ng template DNA, a 100-fold increase over what was typically used for germ line examples. Reproducibility was again improved by using 100ng post-hybridization template DNA, representing a further 20-fold increase in target molecules.

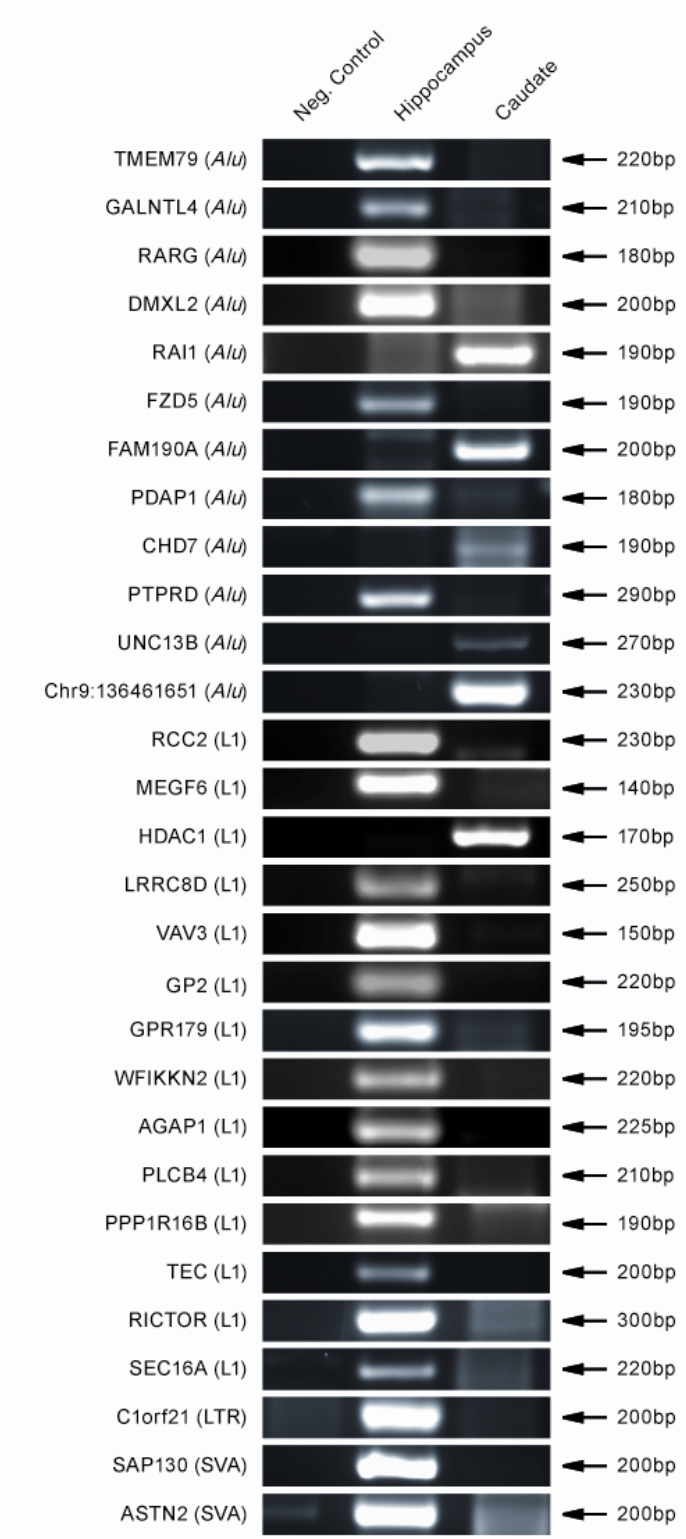
Microarray analysis

Human tissue atlas data were downloaded from the NCBI Gene Expression Omnibus (GSE7905)²⁶. Three replicates of each of the following tissues were included: lung, liver, brain, skeletal muscle, small intestine, skin, pancreas, kidney and spleen. Raw expression values were log₂ normalized to the median. Differential expression was then calculated using the Bioconductor package Limma²⁷ and Benjamini-Hochberg correction for multiple testing. Genes differentially expressed (P<0.05) at least 2-fold higher in brain than other tissues were selected. We then intersected this list with the catalogue of RefSeq genes containing somatic L1 insertions (Supplementary Table S4) and calculated enrichment using a χ^2 test.

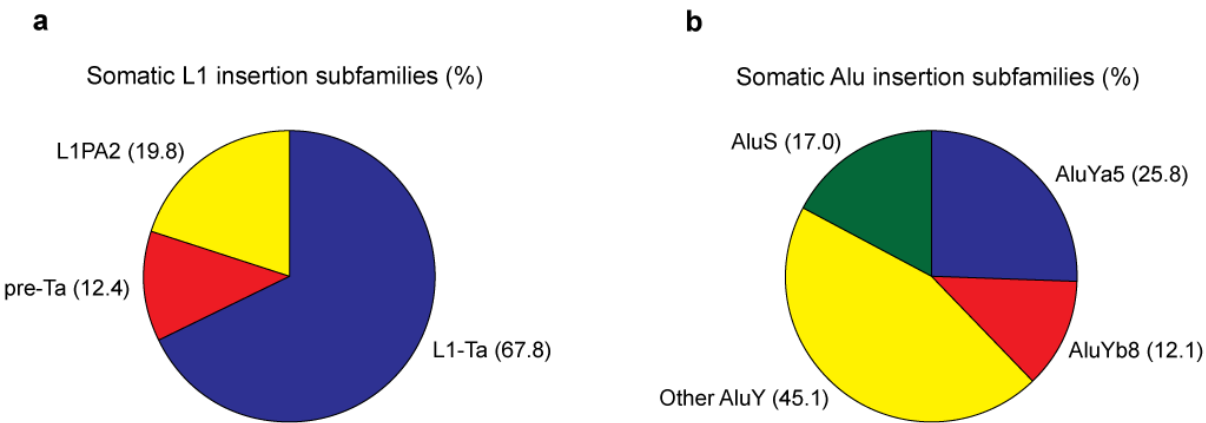
Supplementary Figures



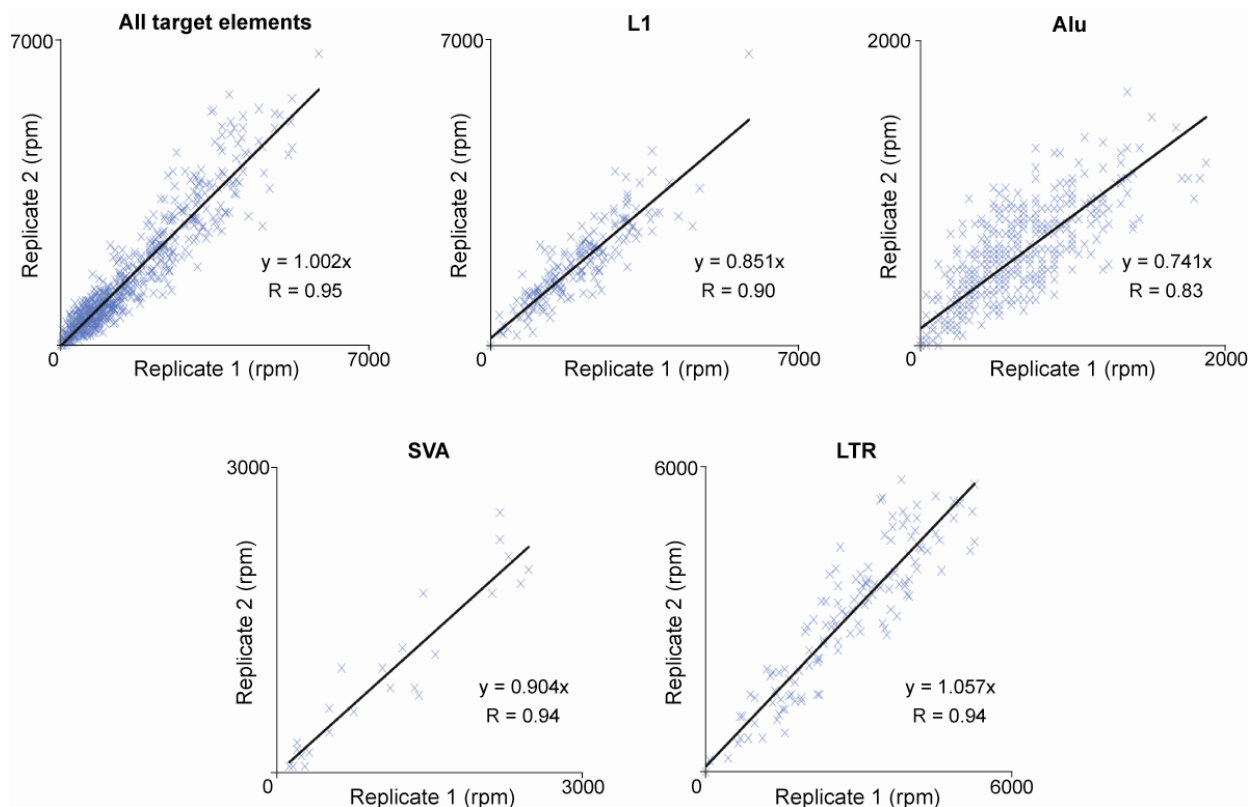
Supplementary Figure S1: Four general types of somatic L1 insertion detected by RC-seq. Capture array probes targeting the termini of L1 sequences yield RC-seq reads that span insertion 5' and 3' junctions (illustrated as RC-seq reads 1-3). These are then mapped to the reference genome and correspond to (a) the 5' end of full-length or near full-length insertions (~10% of junctions detected) or (b) truncated insertions sequenced at their 5' end and captured by a probe at their 3' end (~70% of junctions detected) or (c) the 3' end of insertions of unknown length (~20% of junctions detected) or (d) insertions of known length detected at both ends (rare). As illustrated here, only (d) can conclusively determine whether target site duplications (TSDs) occur and therefore whether retrotransposition is driven by TPRT or an endonuclease independent mechanism. Also note that, as the average RC-seq insert size was ~250nt for most libraries, paired-end 101mer reads were usually separated by a gap of ~50nt (as depicted here). Longer 2x150mer sequencing performed on donor B samples, with an insert size of ~200nt, served to remove this gap.



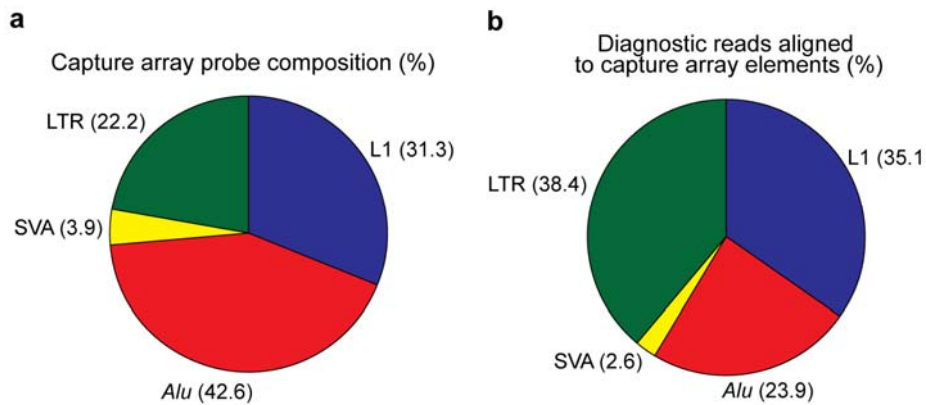
Supplementary Figure S2: PCR confirmation of somatic L1, *Alu*, SVA and LTR insertions. Further details for each insertion are provided in Supplementary Table S7.



Supplementary Figure S3: Subfamily annotations for L1 (a) and *Alu* (b) somatic insertions. n = 7,743 for (a) and n = 13,692 for (b).



Supplementary Figure S4: Quantitative reproducibility of sequencing depth per element based on two RC-seq technical replicates from donor A caudate nucleus. Given the substantially higher sequencing depth of Replicate 2, values were normalized to reads-per-million (rpm) mapped diagnostically to a target element. Pearson correlation (R) and gradient of linear regression ($y = mx$) are indicated in each graph. The number of elements included for each comparison (n) is given in Supplementary Table S1.



Supplementary Figure S5: RC-seq capture and sequencing bias is limited. (a) Proportions of capture array probes by retrotransposon family. (b) Proportions of diagnostic reads aligning to each retrotransposon family across all brain libraries.

Supplementary Tables

Family	Selection criteria (w.r.t. full-length consensus)				Count
	coverage ≥ (%)	divergence ≤ (%)	insertions ≤ (%)	deletions ≤ (%)	
L1	98.0	2.0	1.0	1.0	191
<i>Alu</i>	99.0	0.5	0.0	0.0	520
SVA	95.0	5.0	2.0	2.0	24
LTR	98.0	2.0	1.0	1.0	140

Supplementary Table S1: Selection criteria and counts of retrotransposon families targeted for sequence capture. Minimum coverage and maximum divergence, insertion and deletion percentages were chosen to produce a mixture of probes covering those L1, *Alu* and SVA elements that were most likely to mobilize *in vivo*. ERVK and ERV1 LTRs are unlikely to mobilize in humans¹⁹ and, after fulfilling the same selection criteria as L1, were included as negative controls.

Family	Sampling	Exonic	Intronic	Intergenic
L1	Putative somatic	0.034 (260)	0.433 (3354)	0.533 (4129)
	Known germ line	0.009 (26)	0.321 (956)	0.671 (1999)
	Random expectation	0.022	0.356	0.622
<i>Alu</i>	Putative somatic	0.035 (481)	0.385 (5266)	0.580 (7945)
	Known germ line	0.009 (55)	0.394 (2299)	0.597 (3485)
	Random expectation	0.022	0.356	0.622

Supplementary Table S8: Observed frequencies of L1 and *Alu* insertions in protein-coding genes. Proportions are displayed for exonic, intronic and intergenic insertions, with raw counts in brackets. Known germ line insertion counts were compiled from other sources^{1,16,18}. Random expectation statistics reflect the proportions of exons and introns in the human reference genome based on RefSeq annotations.

Cluster	GO ID	GO Term	% Genes	P-value (BH corrected)
1 (10.7)	GO:0045202	Synapse	9.7	1.3E-11
	GO:0044456	synapse part	7.0	1.1E-08
	GO:0045211	postsynaptic membrane	4.2	1.2E-05
2 (9.4)	GO:0007155	cell adhesion	15.2	2.2E-09
	GO:0022610	biological adhesion	15.2	2.5E-09
	GO:0016337	cell-cell adhesion	5.6	4.5E-01
3 (8.4)	GO:0006468	protein amino acid phosphorylation	13.7	2.1E-06
	GO:0006793	phosphorus metabolic process	18.0	1.9E-05
	GO:0006796	phosphate metabolic process	18.0	1.9E-05
	GO:0016310	phosphorylation	15.2	5.9E-05
4 (7.3)	GO:0030030	cell projection organization	9.2	6.4E-08
	GO:0000904	cell morphogenesis involved in differentiation	6.4	1.5E-05
	GO:0048667	cell morphogenesis involved in neuron differentiation	5.7	2.6E-05
	GO:0032990	cell part morphogenesis	6.4	1.1E-04
	GO:0048858	cell projection morphogenesis	6.2	1.1E-04
	GO:0000902	cell morphogenesis	8.1	1.2E-04
	GO:0048812	neuron projection morphogenesis	5.6	1.4E-04
	GO:0007409	axonogenesis	5.2	2.0E-04
	GO:0032989	cellular component morphogenesis	8.6	4.5E-04
	GO:0006928	cell motion	9.8	6.3E-04
	GO:0031175	neuron projection development	6.1	1.0E-03
	GO:0007411	axon guidance	3.3	3.3E-03
	GO:0030182	neuron differentiation	8.9	4.9E-03
	GO:0048666	neuron development	7.3	5.7E-03

Supplementary Table S9: Gene Ontology (GO) analysis of genes containing intronic somatic L1 mutations. Functional term enrichment and clustering were performed using the database for annotation, visualization and integrated discovery (DAVID)²⁸, default clustering parameters and the BP_FAT and CC_FAT gene ontologies. The top four clusters are listed followed by their enrichment scores in brackets. A parallel analysis conducted using 365 RefSeq genes known to contain germ line L1-Ta and pre-Ta insertions annotated by RepeatMasker revealed a similar, though substantially weaker, enrichment for neurobiological terms.

Family	Detected Rep 1 and Rep 2	Detected Rep 1 not Rep 2	Reproducibility (%)
L1	180	0	100.0
<i>A/u</i>	440	6	98.7
SVA	24	0	100.0
LTR	128	2	98.5
All	772	8	99.0

Supplementary Table S10: RC-seq qualitative reproducibility based on technical replicates from donor A caudate nucleus. Detection was defined as having at least one diagnostic read pair spanning either the 5' or 3' terminus of an element.

Family	Detected 5' and 3' (%)	Detected only 5' or only 3' (%)	Not detected (%)	Average reads per detected element
L1	60.7	33.5	5.8	71.1
<i>Alu</i>	68.8	16.9	14.3	18.2
SVA	95.8	4.2	0.0	41.5
LTR	83.6	9.3	7.1	96.9
All	70.2	18.9	10.9	44.2

Supplementary Table S11: RC-seq coverage and sequencing depth for donor A caudate nucleus Replicate 1.
Statistics are for “diagnostic” reads only.

Supplementary References

1. Wang, J., *et al.* dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat* **27**, 323-329 (2006).
2. Venter, J.C., *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
3. Brouha, B., *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**, 5280-5285 (2003).
4. Bennett, E.A., Coleman, L.E., Tsui, C., Pittard, W.S. & Devine, S.E. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**, 933-951 (2004).
5. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements are active in the human genome? *Trends Genet* **23**, 183-191 (2007).
6. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
7. Coufal, N.G., *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127-1131 (2009).
8. Muotri, A.R., *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903-910 (2005).
9. Muotri, A.R., *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443-446 (2010).
10. Cost, G.J., Golding, A., Schlissel, M.S. & Boeke, J.D. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res* **29**, 573-577 (2001).
11. Bennett, E.A., *et al.* Active Alu retrotransposons in the human genome. *Genome Res* **18**, 1875-1883 (2008).
12. Comeaux, M.S., Roy-Engel, A.M., Hedges, D.J. & Deininger, P.L. Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? *Genome Res* **19**, 545-555 (2009).
13. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, 41-48 (2003).
14. Ostertag, E.M., Goodier, J.L., Zhang, Y. & Kazazian, H.H., Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**, 1444-1451 (2003).
15. Beck, C.R., *et al.* LINE-1 Retrotransposition Activity in Human Genomes. *Cell* **141**, 1159-1170 (2010).
16. Ewing, A.D. & Kazazian, H.H., Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**, 1262-1270 (2010).
17. Huang, C.R., *et al.* Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. *Cell* **141**, 1171-1182 (2010).
18. Iskow, R.C., *et al.* Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. *Cell* **141**, 1253-1261 (2010).
19. Lander, E.S., *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
20. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996).
21. Fu, Y., *et al.* Repeat subtraction-mediated sequence capture from a complex genome. *Plant J* **62**, 898-909 (2010).
22. Faulkner, G.J., *et al.* A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91**, 281-288 (2008).

23. Li, R., *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).
24. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
25. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
26. Dezso, Z., *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol* **6**, 49 (2008).
27. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).
28. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).