Edinburgh Research Explorer

Somatic retrotransposition alters the genetic landscape of the human brain

OPEN ACCESS

Somatic retrotransposition alters the genetic landscape of the human brain

J. Kenneth Baillie[1,*], **Mark W. Barnett**[1,*], **Kyle R. Upton**[1,*], **Daniel J. Gerhardt**[2], **Todd A. Richmond**[2], **Fioravante De Sapio**[1], **Paul Brennan**[3], **Patrizia Rizzu**[4], **Sarah Smith**[1], **Mark Fell**[1], **Richard T. Talbot**[1], **Stefano Gustincich**[5], **Thomas C. Freeman**[1], **John S. Mattick**[6], **David A. Hume**[1], **Peter Heutink**[4], **Piero Carninci**[7], **Jeffrey A. Jeddeloh**[2], and **Geoffrey J. Faulkner**[1,†]

[1]Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, EH25 9RG, UK [2]Roche NimbleGen, Inc., 500 S. Rosa Road, Madison, WI 53719, USA [3]Edinburgh Cancer Research Centre, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XR, UK [4]Section of Medical Genomics, Department of Clinical Genetics, VU University Medical Center, Van der Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands [5]Sector of Neurobiology, International School for Advanced Studies (SISSA), via Bonomea, 265, 34136 Trieste, Italy [6]Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072, Australia [7]RIKEN Yokohama Institute, Omics Science Center, 1-7-22 Suehiro-chô, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Abstract

Retrotransposons are mobile genetic elements that employ a germ line "copy-and-paste" mechanism to spread throughout metazoan genomes[1]. At least 50% of the human genome is derived from retrotransposons, with three active families (L1, *Alu* and SVA) associated with insertional mutagenesis and disease[2-3]. Epigenetic and post-transcriptional suppression block retrotransposition in somatic cells[4-5], excluding early embryo development and some malignancies[6-7]. Recent reports of L1 expression[8-9] and copy number variation[10-11] (CNV) in the human brain suggest L1 mobilization may also occur during later development. However, the corresponding integration sites have not been mapped. Here we apply a high-throughput method to identify numerous L1, *Alu* and SVA germ line mutations, as well as 7,743 putative somatic L1 insertions in the hippocampus and caudate nucleus of three individuals. Surprisingly, we also found 13,692 and 1,350 somatic *Alu* and SVA insertions, respectively. Our results demonstrate that retrotransposons mobilize to protein-coding genes differentially expressed and active in the brain. Thus, somatic genome mosaicism driven by retrotransposition may reshape the genetic circuitry that underpins normal and abnormal neurobiological processes.

Malignancy and ageing are commonly associated with the accumulation of deleterious mutations that lead to loss of function, cell death or uncontrolled growth. Retrotransposition is clearly mutagenic; an estimated 400 million retrotransposon-derived structural variants are present in the global human population[3] and more than 70 diseases involve heritable and *de novo* retrotransposition events[2]. Presumably for this reason transposition-competent retrotransposons are heavily methylated and transcriptionally inactivated[4-5]. Nevertheless, substantial somatic L1 retrotransposition has been detected in neural cell lineages[10-12]. Given the complex structural and functional organization of the mammalian brain, its adaptive and regenerative capabilities[13] and the unresolved etiology of many neurobiological disorders, these somatic insertions could be of major significance[14].

One explanation for the observed transpositional activity in the brain may be that the L1 promoter is transiently released from epigenetic suppression during neurogenesis[11-12]. Transposition-competent L1s can then repeatedly mobilize to different loci in individual cells and produce somatic mosaicism. Several lines of evidence support this model, including L1 transcription[8-9] and CNV in brain tissues from human donors of various ages[10-11] as well as mobilization of engineered L1s *in vitro* and in transgenic rodents[10,12]. Importantly, it is not known where somatic L1 insertions occur in the genome nor, considering that open chromatin is susceptible to L1 integration[15], whether these events disproportionately affect protein-coding loci expressed in the brain.

Mapping the individual retrotransposition events that collectively form a somatic mosaic is challenging due to the rarity of each mutant allele in a heterogeneous cell population. We therefore developed a high-throughput protocol called retrotransposon capture sequencing (RC-seq). Firstly, fragmented genomic DNA was hybridized to custom sequence capture arrays targeting the 5′ and 3′ termini of full-length L1, *Alu* and SVA retrotransposons (Fig . 1a, Supplementary Table S1, Supplementary Table S2). Immobile ERVK and ERV1 LTR elements were included as negative controls. Secondly, the captured DNA was deeply sequenced, yielding ~25 million paired-end 101mer reads per sample (Fig. 1b). Lastly, read pairs were mapped using a conservative computational pipeline designed to identify known (Fig. 1c) and novel (Fig. 1d, Supplementary Fig. S1a-d) retrotransposon insertions with uniquely mapped read pairs ("diagnostic reads") spanning their termini.

Previous works have equated L1 CNV with somatic mobilization *in vivo*[10-11]. To test this assumption with RC-seq, we first screened five brain sub-regions taken from three individuals (donors A, B and C) for L1 CNV. A significant ($p<0.001$) increase was observed in the number of L1 ORF2 copies present in DNA extracted from the hippocampus of donor C and a similar though smaller increase for donor A (Fig. 2). RC-seq was then applied to the brain regions that exhibited the highest (hippocampus) and lowest (caudate nucleus) L1 CNV using samples from all three donors, including a technical replicate of donor A caudate nucleus. A total of 177.4 million RC-seq paired-end reads were generated from seven libraries (Supplementary Table S3). RC-seq achieved deep sequencing coverage of known active retrotransposons, high reproducibility and limited sequence capture bias (Supplementary Results).

Read pairs diagnostic for novel retrotransposon insertions were clustered based on their insertion site, relative orientation and retrotransposon family. A total of 25,229 clusters were produced. Proximal clusters arranged on opposing strands indicated two termini of one insertion and were paired, resulting in a catalogue of 24,540 novel insertions (Supplementary Table S4). Unsurprisingly, the vast majority of these were either L1 (32.2%) or *Alu* (60.9%) (Fig. 3a). To segregate germ line mutations from other events, we combined the three largest available catalogues of L1 and *Alu* polymorphisms[6,16-17] as an annotation database and also performed RC-seq upon genomic DNA extracted from pooled

human blood, producing 6,150 clusters (Supplementary Table S5) that were intersected with the existing brain RC-seq clusters. Any brain clusters that (a) contained RC-seq reads from more than one region or individual, (b) overlapped a blood RC-seq cluster or (c) matched a known polymorphism were designated as germ line insertions. Overall, 8.4% of *Alu* insertions in the brain were annotated as germ line, versus only 1.9% for L1. Nearly all unannotated L1 insertions matched fewer than three diagnostic RC-seq reads (Fig. 3b) and were considered potential somatic insertions.

Candidate insertions were validated by PCR amplification and capillary sequencing. Thirty-five germ line L1, *Alu*, SVA and LTR insertions readily confirmed by single-step PCR (Supplementary Table S6). Given low target molecule abundance and the high genomic frequency of the L1 3′ end, we devised a 5′ end nested PCR validation assay for somatic insertions. From 850 and 2,601 full-length (90%) L1 and *Alu* insertions, respectively representing 11.0% and 19.0% of the putative somatic insertions found for each family, we selected 29 examples (14 L1, 15 *Alu*) for validation. Nearly all of the chosen examples were exonic or intronic and were prioritized based on the degree of 5′ truncation, with longer insertions preferred. Optimization of the protocol, combined with substantial input DNA (100ng) ultimately lead to the confirmation of 14/14 L1s and 12/15 *Alu*s (Supplementary Table S7, Supplementary Fig. S2). Four somatic SVA insertions were also assayed using the same process and two confirmed (both SVA_F) before the available input material was exhausted.

Repeated attempts to PCR amplify the corresponding 3′ junctions consistently yielded off-target amplicons, leaving validation based exclusively on 5′ junctions. For this reason, we could not experimentally identify the target site duplications (TSDs) that are indicative of retrotransposition via target-primed reverse transcription (TPRT)[1]. We propose that the 3′ junctions of insertions validated at their 5′ end did not amplify efficiently due to the confounding factors listed above, as well as the presence of long polyA tails in on-target amplicons but often not, as we found, in off-target amplicons.

However, TSDs could in some cases be found directly by RC-seq (Supplementary Fig. 1d). An examination of germ line insertions sequenced to high depth (10 reads) at both their 5′ and 3′ ends revealed that 43/50 (86%) presented TSDs. Due to their very low abundance - and therefore low sequencing coverage - only three putative somatic insertions were detected by at least one RC-seq read at both termini. Two of these examples (one L1 and one *Alu*) presented TSDs. Despite these and other data strongly supporting retrotransposition as the main cause of somatic mobilization (Supplementary Results) an insufficient number of examples were sequenced at both ends to distinguish whether TPRT or an alternative retrotransposition mechanism[18] was primarily responsible.

The somatic origin of each insertion was demonstrated by its presence in one of the assayed brain tissues and absence from the other, according to RC-seq and PCR results. As illustrative examples, an intronic somatic L1 insertion in HDAC1 is detailed in Fig. 4a and Fig. 4b whilst an exonic somatic *Alu* insertion in RAI1 is shown in Fig. 4c and Fig. 4d. These experimental results indicated that insertions detected by RC-seq occurred *in vivo* and did not represent sequencing artifacts.

Donor element annotation revealed that 80.2% of somatic L1 insertions corresponded to the most recently active human L1 subfamilies, L1-Ta and pre-Ta (Supplementary Fig. S3a). The normalized hippocampus:caudate nucleus ratio for somatic L1 insertions was 1.3, 0.5 and 2.2 for donors A, B and C, respectively, paralleling trends from the L1 CNV assay (Fig. 2). Protein-coding loci were disproportionately affected (Supplementary Table S8) compared to random expectation and compared to prior germ line frequencies (P<0.0001 for

exons and introns, χ^2 test). Pre-existing microarray expression data indicated that genes containing intronic L1s were twice as likely to be differentially over-expressed in the brain compared to random expectation (P<0.0001, χ^2 test). Key loci were found to contain somatic L1 insertions, including tumor suppressor genes deleted in neuroblastoma and glioma (e.g. CAMTA1), dopamine receptors (e.g. DRD3) and neurotransmitter transporters (SLC6A5, SLC6A6, SLC6A9). Globally, a Gene Ontology analysis revealed enrichment for terms relevant to neurogenesis and synaptic function (Supplementary Table S9).

Unlike L1, *Alu* retrotransposition has not previously been reported in normal brain cells, a major finding of the present work. However, the L1 transposition machinery is known to mobilize *Alu* in trans[19] and 83.0% of the somatic *Alu* insertions corresponded to the AluY subfamily most active in the human germ line (Supplementary Fig. S3b), making the coincidence of somatic L1 and *Alu* mobilization plausible. On a per element basis the observed *Alu* activity was approximately twenty-fold lower than L1 (Supplementary Results). Thus, it is unlikely that *Alu* CNV would be statistically significant if assayed by TaqMan qPCR[10]. The genomic patterns of *Alu* and L1 insertions were also different; somatic *Alu* insertions were not overrepresented in introns but were even more common in exons than L1 (Supplementary Table S8). *Alu* exonization is a noted cause of genetic disease[2]. Overall, L1, *Alu* and, to a more limited extent, SVA mobilization produced a large number of insertions that affected protein-coding genes.

Our results provide clear evidence that somatic L1 and *Alu* mobilization fundamentally alters the genetic landscape of the human brain and that retrotransposition is the primary mechanism underlying this phenomenon. In contrast to germ line activity[6,16], somatic insertions disproportionately impacted protein-coding loci. Germ line insertions are rarely found in regions where they generate a deleterious phenotype because such mutations are strongly selected against during evolution. Somatic events, on the other hand, are present for one generation and may affect protein-coding loci in a specific environmental context, perhaps drawn to open chromatin in transcribed regions[15]. Apart from the obvious effects of exonic insertions, intronic events could act as subtle "transcriptional rheostats"[20] or as cis-regulatory elements[21] akin to the IAP insertion responsible for the viable yellow allele of Agouti in the mouse[22].

Several recent studies have catalogued retrotransposon insertions in the human germ line and tumors[6,16,23-24]. Through RC-seq we have extended these data to the brain and linked somatic retrotransposition to neurobiological genes. For instance, HDAC1 is a genome-wide transcriptional regulator that controls the canonical L1 promoter[4,25] and is implicated in psychiatric disease and tumorigenesis[26]. Another example highlighted here, RAI1, is a transcription factor highly expressed in the brain and previously linked with schizophrenia and Smith-Magenis syndrome[27]. An exonic *Alu* insertion in RAI1, as shown in Fig. 4c, could therefore have phenotypic consequences.

The hippocampus appears predisposed to somatic L1 retrotransposition[10], which is intriguing given that its subgranular zone is a major source of adult neurogenesis[13]. This is also consistent with the hypothesis that L1 retrotransposition is related to neural plasticity[14]. Even more intriguing is the possibility that the APOBECs, RNA/DNA editing enzymes that have expanded under strong positive selection in the primate lineage and been shown to control L1 mobility, may modulate somatic retrotransposition in the brain[28].

Mutagenesis due to somatic retrotransposition has obvious tumorigenic potential[29] and may play a role in other diseases and biological processes. For example, deletion of the chromatin remodeling HDAC1 cofactor MeCP2[12,25] leads to increased L1 copy number and may inhibit neuronal maturation in Rett syndrome[30]. Somatic mosaicism could also be a

factor in neurological dimorphisms seen among discordant monozygotic twins[14]. Future studies may determine whether the overall frequency of somatic retrotransposition varies considerably between individuals, as suggested by our data and previous experiments[10], and between populations. Ultimately, direct identification of transcripts disrupted by somatic retrotransposition, together with its epigenetic regulation, may provide insights into the molecular processes underlying human cognition, neurodevelopmental disorders and neoplastic transformation.

METHODS SUMMARY

Human DNA samples

Tissues were provided by the Netherlands Brain Bank (Amsterdam, The Netherlands) for three post mortem donors without evidence of neurodegeneration. Pooled human genomic DNA was purchased from Promega.

TaqMan qPCR

Quantitative PCR experiments were performed with minor modifications to an earlier approach[10]. Quantification included five technical replicates. For each assay, the ratio of L1 ORF2 to α-satellite repeats (SATA) was normalized to the ratio obtained for caudate nucleus. Ratios were compared across brain regions with a repeated measures one-way ANOVA with Bonferroni correction.

Retrotransposon capture array design

A NimbleGen Sequence Capture 2.1M Array was customized to contain oligonucleotide probes tiled across the termini of full-length L1, *Alu* and SVA retrotransposons, as well as LTRs intended to act as negative controls. Probes were not filtered for repetitiveness. Eight probes were typically generated per L1, SVA and LTR and four probes per *Alu*, with a total of 4,885 probes across 875 targeted elements.

Capture library preparation and sequencing

DNA sequencing libraries were constructed using an Illumina paired-end kit, with substantial modifications (see Supplementary Methods). 2.5μg starting genomic DNA was used for each RC-seq library. Ligation mediated PCR (LM-PCR) based amplification was performed pre and post hybridization. The average insert size was ~250nt. Enrichment was confirmed by qPCR against *Alu*. Sequencing was performed by ARK-Genomics, The Roslin Institute, on an Illumina GAIIx instrument.

Computational analyses

Paired-end RC-seq reads were mapped to hg19 using SOAP2. Reads where both ends could be aligned to the genome, but not at the same locus, indicated novel retrotransposon insertions. These alignments were corroborated by BLAT, stringently filtered and clustered. Clusters were annotated using published retrotransposon databases[6,16-17] and the NCBI RefSeq database.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

REFERENCES

1. Kazazian HH Jr. Mobile elements: drivers of genome evolution. Science. 2004; 303:1626–32. [PubMed: 15016989]

2. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009; 10:691–703. [PubMed: 19763152]

3. Xing J, et al. Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 2009; 19:1516–26. [PubMed: 19439515]

4. Garcia-Perez JL, et al. Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature. 2010; 466:769–73. [PubMed: 20686575]

5. Yang N, Kazazian HH Jr. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol. 2006; 13:763–71. [PubMed: 16936727]

6. Iskow RC, et al. Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. Cell. 2010; 141:1253–1261. [PubMed: 20603005]

7. Kano H, et al. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. 2009; 23:1303–12. [PubMed: 19487571]

8. Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P. Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res. 2010

9. Faulkner GJ, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009; 41:563–71. [PubMed: 19377475]

10. Coufal NG, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009; 460:1127–31. [PubMed: 19657334]

11. Muotri AR, et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010; 468:443–446. [PubMed: 21085180]

12. Muotri AR, et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005; 435:903–10. [PubMed: 15959507]

13. Eriksson PS, et al. Neurogenesis in the adult human hippocampus. Nat Med. 1998; 4:1313–7. [PubMed: 9809557]

14. Singer T, McConnell MJ, Marchetto MC, Coufal NG, Gage FH. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci. 2010; 33:345–54. [PubMed: 20471112]

15. Cost GJ, Golding A, Schlissel MS, Boeke JD. Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res. 2001; 29:573–7. [PubMed: 11139628]

16. Ewing AD, Kazazian HH Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010; 20:1262–70. [PubMed: 20488934]

17. Wang J, et al. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 2006; 27:323–9. [PubMed: 16511833]

18. Morrish TA, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2002; 31:159–65. [PubMed: 12006980]

19. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 2003; 35:41–8. [PubMed: 12897783]

20. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature. 2004; 429:268–74. [PubMed: 15152245]

21. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008; 9:397–405. [PubMed: 18368054]

22. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet. 1999; 23:314–8. [PubMed: 10545949]

23. Beck CR, et al. LINE-1 Retrotransposition Activity in Human Genomes. Cell. 2010; 141:1159–1170. [PubMed: 20602998]

24. Huang CR, et al. Mobile Interspersed Repeats Are Major Structural Variants in the Human Genome. Cell. 2010; 141:1171–1182. [PubMed: 20602999]

25. Nan X, et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature. 1998; 393:386–9. [PubMed: 9620804]

26. Kazantsev AG, Thompson LM. Therapeutic application of histone deacetylase inhibitors for central nervous system disorders. Nat Rev Drug Discov. 2008; 7:854–68. [PubMed: 18827828]

27. Slager RE, Newton TL, Vlangos CN, Finucane B, Elsea SH. Mutations in RAI1 associated with Smith-Magenis syndrome. Nat Genet. 2003; 33:466–8. [PubMed: 12652298]

28. Mattick JS. RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. Bioessays. 2010; 32:548–52. [PubMed: 20544741]

29. Miki Y, et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 1992; 52:643–5. [PubMed: 1310068]

30. Chahrour M, Zoghbi HY. The story of Rett syndrome: from clinic to neurobiology. Neuron. 2007; 56:422–37. [PubMed: 17988628]
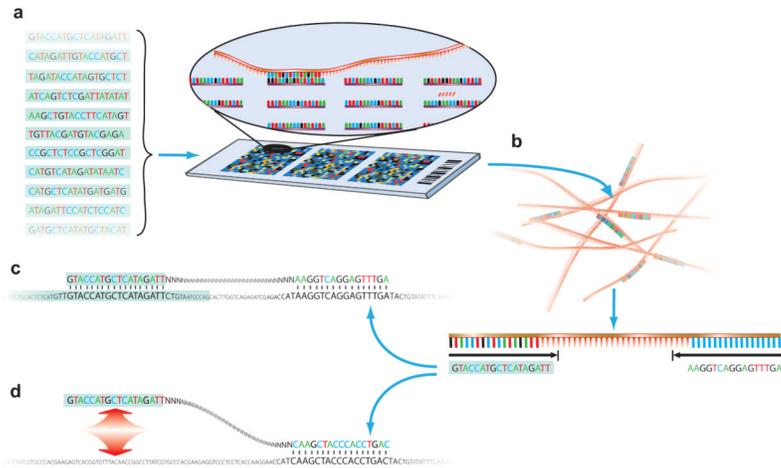
Figure 1.
Overall RC-seq methodology
(a) Retrotransposon capture: sheared genomic DNA is hybridized to custom tiling arrays probing full-length retrotransposons (nucleotides highlighted with light blue background).
(b) Sequencing: post-hybridization, DNA fragments are eluted and analyzed with an Illumina sequencer, producing $\sim 2.5 \times 10^7$ paired-end reads per library that are subsequently aligned to the reference genome. **(c)** Reads mapping as a pair to a single locus indicate known retrotransposon insertions. **(d)** Unpaired reads where one end maps to a single locus and the other end maps to a distal retrotransposon indicate novel retrotransposition events.
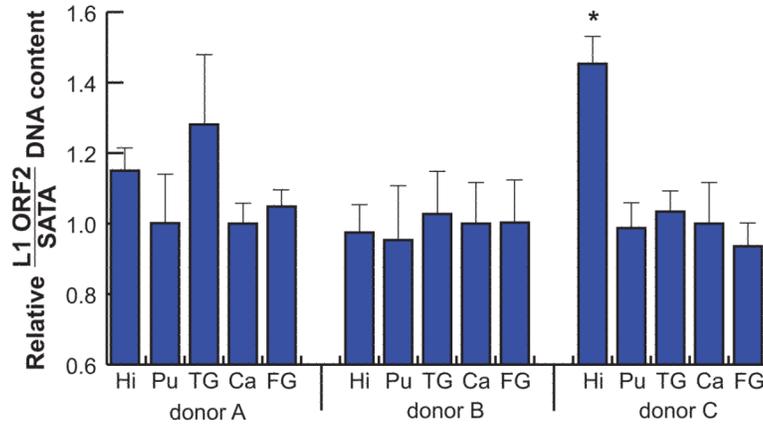
Figure 2. Multiplex qPCR confirms L1 copy number variation in the human brain
The relative abundance of L1 open reading frame 2 (ORF2) versus α-satellite repeats (SATA) was quantified using an existing TaqMan based approach[10]. Genomic DNA from five brain regions was assayed in three donors (A, B and C). Hi, hippocampus; Pu, putamen; TG, middle temporal gyrus; Ca, caudate nucleus; FG, middle frontal gyrus. Values are normalized to caudate nucleus for each donor. Error bars equal one s.e.m. *$p < 0.001$ for repeated measures one way analysis of variance (ANOVA) within each donor, followed by pairwise least significant difference (LSD) post hoc tests with Bonferroni correction.
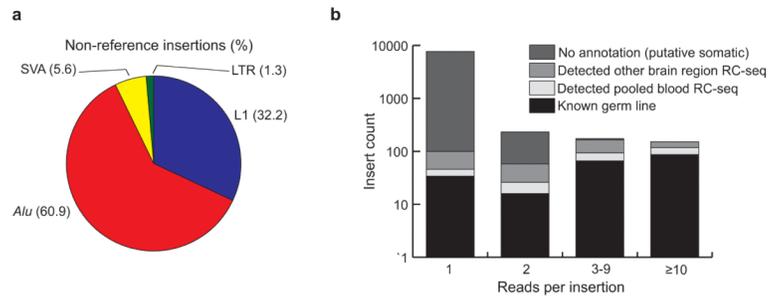
Figure 3. Characterization of non-reference genome insertions
(a) Proportions of novel insertions found for each family. **(b)** Annotation of novel L1 insertions across all brain libraries. The vast majority of insertions detected by fewer than three reads could not be annotated and were considered putative somatic events. Note: logarithmic scale.
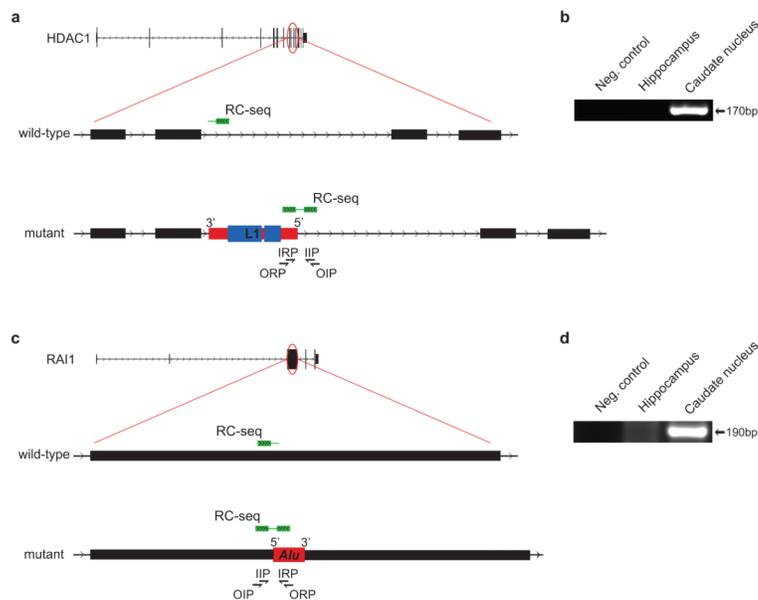
Figure 4. Discovery of somatic insertions in HDAC1 and RAI1
(a) Alignment of an RC-seq read from donor C caudate nucleus indicated an antisense L1 insertion in intron 9 of HDAC1. Nested PCR primers were designed to span the L1 5′ terminus, with an initial reaction combining outside retrotransposon (ORP) and insertion site (OIP) primers and a second reaction using inside retrotransposon (IRP) and insertion site (IIP) primers. **(b)** Amplification of the nested PCR target, confirmed for specificity by capillary sequencing, was achieved in caudate nucleus but not in hippocampus. Sequencing indicated that the L1 mobilized from chromosome 9 and was accompanied by 5′ transduction. **(c)** Alignment of an RC-seq read pair from donor A caudate nucleus indicated a sense *Alu* insertion in exon 3, and the CDS, of RAI1. **(d)** As for (b) amplification of the nested PCR target was achieved in caudate nucleus but not in hippocampus. Sequencing indicated that the *Alu* mobilized from chromosome 4. Note: L1 and *Alu* elements in (a) and (c) are not drawn to scale.