RESEARCH ARTICLE

Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals

Boxun Zhao[1,2], Qixi Wu[3], Adam Yongxin Ye[4,5,6], Jing Guo[1,7], Xianing Zheng[1,2], Xiaoxu Yang[5], Linlin Yan[5], Qing-Rong Liu[8], Thomas M. Hyde[9,10], Liping Wei[1,2,5]*, August Yue Huang[5]*

1 National Institute of Biological Sciences, Beijing, China, **2** Graduate School of Peking Union Medical College, Beijing, China, **3** School of Life Sciences, Peking University, Beijing, China, **4** Peking-Tsinghua Center for Life Sciences, Beijing, China, **5** Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China, **6** Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China, **7** College of Life Sciences, Beijing Normal University, Beijing, China, **8** Laboratory of Clinical Investigation, National Institute on Aging, Baltimore, Maryland, United States of America, **9** Lieber Institute for Brain Development, Baltimore, Maryland, United States of America, **10** Departments of Psychiatry & Behavioral Sciences and Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America

* weilp@mail.cbi.pku.edu.cn (LW); huangy@mail.cbi.pku.edu.cn (AYH)

Citation: Zhao B, Wu Q, Ye AY, Guo J, Zheng X, Yang X, et al. (2019) Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. PLoS Genet 15(4): e1008043. https://doi.org/10.1371/journal.pgen.1008043

Editor: John M. Greally, Albert Einstein College of Medicine, UNITED STATES

Received: August 15, 2018

Accepted: February 21, 2019

Published: April 11, 2019

Data Availability Statement: Sequencing data are available at the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA), accession number SRP114393. All the numerical data for the figures and statistics are provided in Supporting Information.

Funding: This study was supported by the National Natural Science Foundation of China (31530092) and the Ministry of Science and Technology 863 Grant (2015AA020108). The funders had no role in

Abstract

Mounting evidence supports that LINE-1 (L1) retrotransposition can occur postzygotically in healthy and diseased human tissues, contributing to genomic mosaicism in the brain and other somatic tissues of an individual. However, the genomic distribution of somatic human-specific LINE-1 (L1Hs) insertions and their potential impact on carrier cells remain unclear. Here, using a PCR-based targeted bulk sequencing approach, we profiled 9,181 somatic insertions from 20 postmortem tissues from five Rett patients and their matched healthy controls. We identified and validated somatic L1Hs insertions in both cortical neurons and non-brain tissues. In Rett patients, somatic insertions were significantly depleted in exons—mainly contributed by long genes—than healthy controls, implying that cells carrying *MECP2* mutations might be defenseless against a second exonic L1Hs insertion. We observed a significant increase of somatic L1Hs insertions in the brain compared with non-brain tissues from the same individual. Compared to germline insertions, somatic insertions were less sense-depleted to transcripts, indicating that they underwent weaker selective pressure on the orientation of insertion. Our observations demonstrate that somatic L1Hs insertions contribute to genomic diversity and MeCP2 dysfunction alters their genomic patterns in Rett patients.

Author summary

Human-specific LINE-1 (L1Hs) is the most active autonomous retrotransposon family in the human genome. Mounting evidence supports that L1Hs retrotransposition occurs postzygotically in the human brain cells, contributing to neuronal genomic diversity, but

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

the extent of L1Hs-driven mosaicism in the brain is debated. In this study, we profiled genome-wide L1Hs insertions among 20 postmortem tissues from Rett patients and matched controls. We identified and validated somatic L1Hs insertions in both cortical neurons and non-brain tissues, with a higher jumping activity in the brain. We further found that MeCP2 dysfunction might alter the genomic pattern of somatic L1Hs in Rett patients.

Introduction

The term "somatic mosaicism" describes the genomic variations that occur in the somatic cells that make up the body of an individual. These variations contribute to intra-individual genetic diversity among different cells [1]. In addition to various types of cancers, somatic mosaicisms reportedly contribute to a variety of neurological disorders, including epilepsy, neurodegeneration, and hemimegalencephaly [2]. The human-specific LINE-1 (L1Hs) retrotransposon family is the only known family of active autonomous transposons in the human genome [3, 4]. L1s retrotranspose through a process called target-primed reverse transcription (TPRT), with the capacity for *de novo* insertion into new genomic locations in both germline and somatic cells [5, 6]. Mounting evidence supports that L1Hs elements, with increased copy number in the brain relative to other tissues, contribute to neuronal diversity via somatic retrotransposition [7–12].

Recent studies reported the occurrence of somatic L1Hs insertions during neurogenesis and in non-dividing mature neurons [9, 13]. Other studies have observed dysregulated L1Hs copy number in patients with Rett syndrome [8] and schizophrenia [14]. Methyl-CpG binding protein 2 (*MECP2*) is the major disease-causing gene of Rett syndrome [15]. Its gene product, MeCP2, can bind to the 5' UTR of L1 elements and represses their expression and retrotransposition [16]. While it is known that L1 expression and copy number are elevated in the brains of *Mecp2* knockout mice as well as in patients with Rett syndrome [8, 17], little is known about the genomic distribution patterns of somatic L1Hs insertions in Rett patients and healthy individuals.

In contrast to germline insertions, the effects of somatic transposon insertions depend not only on their genomic location. Rather, the specific timing, tissue, and cell lineage at which they occur profoundly influence the impact of somatic insertions [18]. Single-cell targeted sequencing approaches have been used to identify somatic insertions [7, 11, 12]. However, such methods typically require a large number of cells and demand considerable sequencing depth for unbiased profiling of human tissues [19, 20]. Furthermore, owing to the rarity of somatic insertions, investigations of the clonal diversity of somatic insertions would require the sequencing of even larger numbers of cells [19]. Another limitation of single-cell sequencing approaches is that errors of allelic dropout and locus dropout, which frequently occur during the whole genome amplification (WGA) step of library construction, can reduce the sensitivity and specificity of somatic insertion detection. Estimates of the rate of somatic L1Hs insertions vary widely in single-cell genomics studies [21]. Bulk sequencing approach can potentially overcome these limitations and enable the genome-wide identification and quantification of somatic L1Hs insertions, but their low allele frequency in cell populations poses a great challenge to distinguishing true insertion events from technical artifacts [22].

Here, we introduced a PCR-based multiplex bulk sequencing method for sensitive enrichment and specific identification of L1Hs insertions from various types of human tissues. We used this method to perform genome-wide L1Hs insertion profiling of 20 postmortem tissues from five patients with Rett syndrome and their matched healthy controls. The aims of this study were to explore the genomic patterns of somatic L1Hs insertions in neuronal and non-neuronal samples, and to investigate whether MeCP2 dysfunction could alter the distribution of L1Hs retrotransposition in patients with Rett syndrome.

Results

A bulk sequencing method to identify L1Hs insertions

Systematic genome-wide profiling of somatic L1Hs insertions requires effective enrichment of insertion signals and specific identification of true signals from background noise. Enriching neuronal nuclei from bulk brain tissue facilitates the accurate deciphering of cell type-specific characteristic and increases the chance of identifying clonal somatic insertions that are derived from the same progenitor cell and shared by multiple neurons. Therefore, we labeled prefrontal cortex (PFC) neuronal nuclei using an antibody against neuron-specific marker NeuN [23], and subsequently purified NeuN^{+} nuclei from postmortem human PFC by fluorescence-activated cell sorting (FACS) (Fig 1A; S1A–S1D Fig; S1 Appendix). All initially sorted nuclei were re-analyzed with a second round of FACS, and the purity of the initial sorting was found to be $> 96\%$ (S1E and S1F Fig; S1 Appendix). The integrity and purity of sorted nuclei were confirmed by fluorescence microscopy (S2A–S2C Fig).

To distinguish the signals of active L1Hs elements from other transposon families that are typically inactive in human, we developed a method called human active transposon sequencing (HAT-seq) (Fig 1B; S3A Fig; S1 Table) based on ATLAS [24] and several versions of high-throughput sequencing-based L1 amplification methods [7, 25–27]. Firstly, L1Hs insertions were specifically enriched and amplified using a primer targeting the diagnostic "AC" motif of L1Hs [3, 28]. To ameliorate the poor performance of Illumina sequencing platform for low-diversity libraries, we employed a nucleotides-shifting design by adding two, four, or six random nucleotides upstream of the L1Hs-specific primer, which greatly increased the diversity of the structure-transformed semi-amplicon library and markedly improved the sequencing quality of L1Hs 3' end. The constructed libraries preserved information regarding the insertion direction and were sequenced by multiplexed 150 bp paired-end reads. This approach provided sequence information fully spanning the 3' L1Hs-genome junction of each of L1Hs insertions, which enabled the identification of integration sites and facilitated *in silico* false-positive filtering based on both sequence features and read-count.

Genomic position of each L1Hs insertion was determined by the alignment of its 3' flanking sequence (Fig 1C). A custom data analysis pipeline classified putative insertions into one of the following four categories: known reference (KR) germline insertions, known non-reference (KNR) germline insertions, unknown (UNK) germline insertions, and putative somatic insertions (S3B Fig). To further remove technical artifacts induced by non-specific or chimeric amplification and read misalignment in next-generation sequencing, we designed a series of stringent error filters to remove them in different aspects (Table 1): 1) read pairs with non-specific amplification signals and incorrect 3' truncation were removed based on the sequence of L1Hs 3' end (Read 2); 2) after merging paired-end reads into contigs, chimeric molecules with abnormal contig structures were identified by BLAST and filtered out; 3) reads with inconsistencies in BWA-MEM and BLAT alignments were defined as mapping errors; and 4) putative somatic insertion signals without multiple PCR duplicates or those present in different individuals were removed, as they were deemed likely to have resulted from sequencing errors. After applying these error filters, the remaining insertions were annotated with peak features to facilitate downstream analysis.

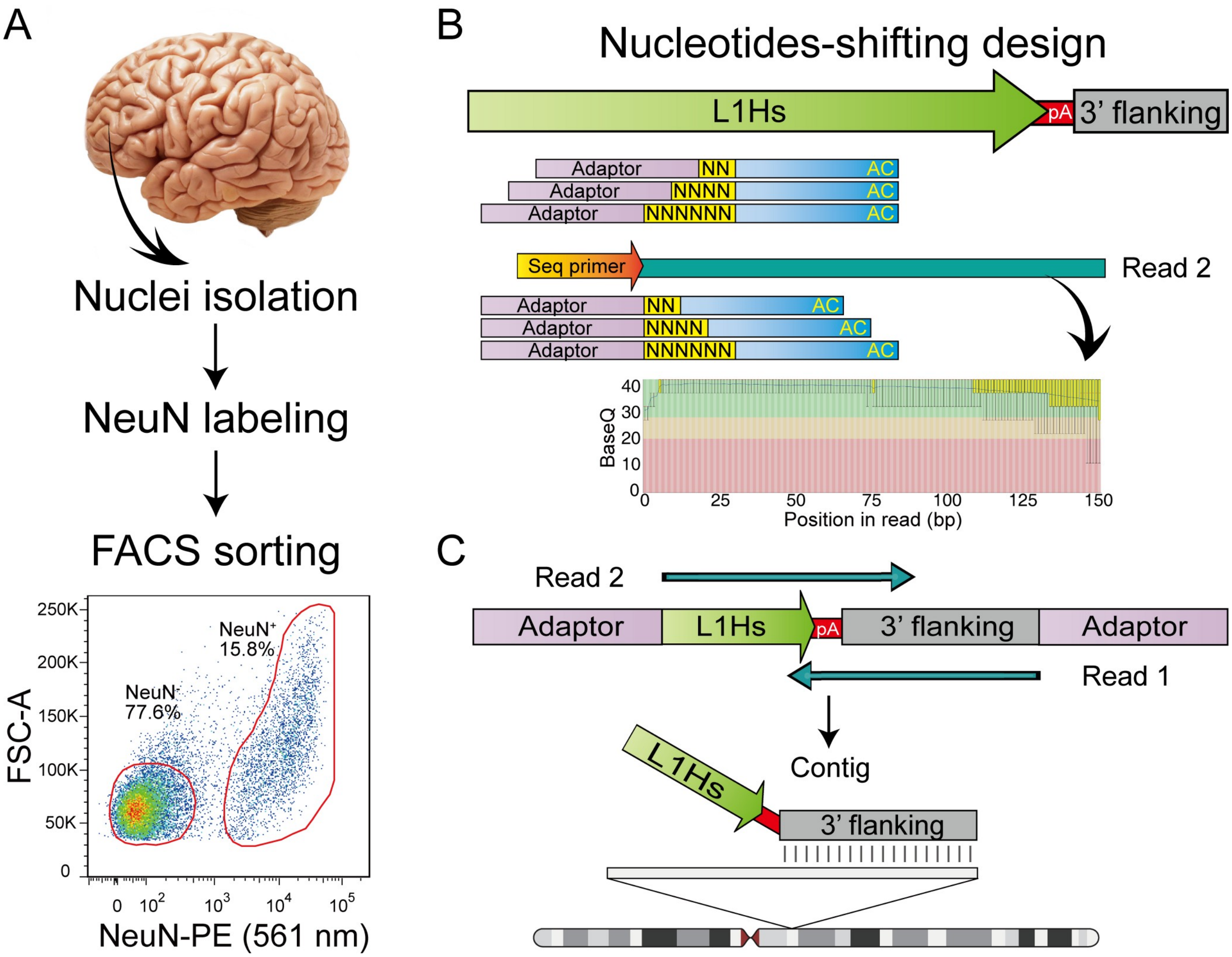

Fig 1. Overview of human active transposon sequencing (HAT-seq). (A) Fluorescence-activated cell sorting (FACS) of prefrontal cortex (PFC) nuclei labeled with NeuN. Two populations (NeuN$^+$ and NeuN$^-$) were sorted. (B) Schematic of the nucleotides-shifting design of the HAT-seq method. By adding two, four, or six random nucleotides upstream of L1Hs-specific primer (L1Hs-AC-28), we transformed the library from a uniform phase-0 amplicon library to a mixed library with phase-2, phase-4, and phase-6 amplicons, which remarkedly improved the base calling accuracy in Read 2. (C) HAT-seq libraries were sequenced with paired-end 150-bp reads. After merging paired reads into contigs that fully spanned the L1Hs-genome 3' junction, genomic locations of each L1Hs insertion were determined by the alignments of their 3' flanking genomic sequences. NeuN-PE, PE-conjugated anti-NeuN antibody.

https://doi.org/10.1371/journal.pgen.1008043.g001

Performance evaluation of the HAT-seq method using a positive control

To benchmark the performance of HAT-seq for detecting somatic L1Hs insertions, we experimentally generated a series of positive control samples with insertions at different frequencies by mixing the genomic DNA (gDNA) extracted from the blood samples of two unrelated adults, ACC1 and ACC2 (see details in Methods). 172 ACC1 non-reference germline L1Hs insertions were identified by HAT-seq, 64 of which were confirmed to be ACC1-specific by 3' junction PCR (3' PCR) analysis of gDNA from ACC1 and ACC2 (Fig 2A; S2 Table; S2 Appendix) and thus served as positive controls. Three HAT-seq libraries were generated from samples consisting of ACC2 gDNA spiked with 1%, 0.1%, or 0.01% of ACC1 gDNA. Considering that decreasing the number of cells pooled for sequencing increased the signal-to-noise ratio for detecting somatic insertions [22], each HAT-seq library was constructed from 20 ng input (about 3,000 cells).

Table 1. Error filters used in the computational pipeline.

Filter name	Definition
Improper alignment	We rejected reads with less than 30-bp alignment or more than 3 mismatches to the reference genome.
Diagnostic motif	We rejected reads without L1Hs diagnostic G motif (position 6012 relative to the L1Hs Repbase consensus) [29].
Chimera within L1 segment	We rejected reads with less than 95% identity (> 4 mismatches) to the L1Hs 3' end consensus sequence.
Chimera within poly-A tail	We rejected reads at risk of being chimeric [12]. We applied BLAST to find the best alignments for retrotransposon and non-retrotransposon segments from hg19. Reads were removed as a putative chimera when the sequences of two segments overlapped > 10 bp with A% (percentage of adenine nucleotides) ≥ 50% or overlapped 6–10 bp with A% < 50%.
Subfamily filter	We rejected putative somatic insertion sites that overlapped with L1 young subfamilies (L1Hs and L1PA2–4) reference insertions.
Known non-reference filter	We rejected putative somatic insertion sites that overlapped with known non-reference L1 insertions in euL1db [30].
Misaligned reads	We rejected reads at risk of being misaligned, defined as inconsistent BWA and BLAT alignment.
Local structural variation (SV)	We rejected reads at risk of being derived from a nearby reference L1Hs [12]. We extracted 2 kb downstream of aligned non-retrotransposon segment from hg19 and aligned the full contig against this sequence by BLAT to exclude potential genomic rearrangement events.
Observed in common	We rejected putative somatic insertion sites observed in two or more individuals.
PCR duplicate	We rejected somatic insertion sites without supporting PCR duplicates.

https://doi.org/10.1371/journal.pgen.1008043.t001

The zygosity of ACC1-specific L1Hs insertions was confirmed by full-length PCR: 49 of which were heterozygous, 9 of which were homozygous, and 6 of which were zygosity-undetermined (Fig 2B; S2 Table; S2 Appendix). We detected all 64 ACC1-specific insertions in our positive control 1% ACC1 spike-in library, 49 (76.6%) of which passed all of error filters and subsequently were deemed "identified" by HAT-seq. In the 0.1% library, we detected 23 ACC1-specific insertions (16 heterozygous, 4 homozygous, and 3 zygosity-undetermined), 17 (73.9%) of which were identified. In the 0.01% library, we detected seven heterozygous ACC1-specific insertions, five (71.4%) of which were identified. The distributions of signal counts (reads with unique start positions) per ACC1-specific insertion followed the Poisson distribution (Fig 2C), indicating a similar probability for each of ACC1-specific insertions to be randomly sampled. In the 1%, 0.1%, and 0.01% libraries, each of ACC1-specific insertions was diluted to 30, 3, and 0.3 copies. Theoretically, by Poisson statistics, there would be 64, 60.81, and 16.59 ACC1-specific insertions being sampled and subsequently being used as the input of HAT-seq libraries (see details in Methods). Therefore, we estimated the sensitivity of HAT-seq for somatic L1Hs insertions in 1%, 0.1%, and 0.01% libraries as 76.6% (49/64), 28% (17/60.81), and 30.1% (5/16.59), respectively. Our data showed that, with about 3,000 cells as input, HAT-seq was able to detect somatic insertion events present in a single cell (Fig 2D and S3 Appendix).

To further evaluate the efficacy of our L1Hs identification pipeline, we compared the proportions of true positives and false positives after applying all the error filters. For the most stringent evaluation, only those 64 ACC1-specific germline insertions in spike-in libraries were defined as "true positives"; all other signals were defined as "false positives", which might include both background noise and some true somatic insertions present in the blood gDNA. As shown in Fig 2E, in three positive control experiments with 1%, 0.1%, and 0.01% ACC1 gDNA spike-in, 76.56%, 73.91%, and 71.43% of true positives remained after all filters, whereas

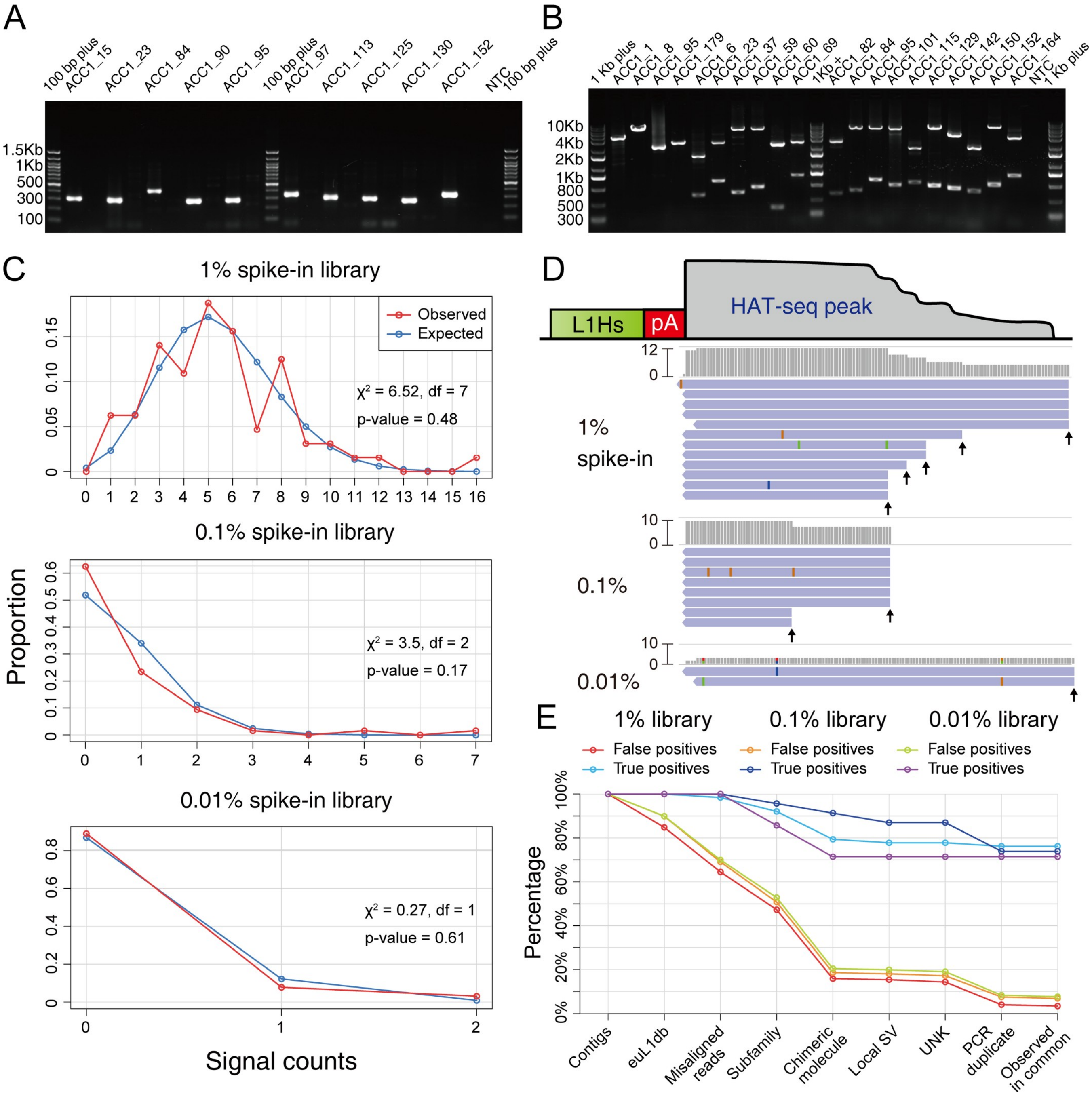

Fig 2. HAT-seq performance evaluation using a positive control. (A) Representative gel image used for the identification of ACC1-specific insertions based on 3' PCR analysis. For each site, genomic DNA from ACC1 and ACC2 was amplified using the same protocol and the PCR products were run on the gel side-by-side (left: ACC1; right: ACC2). NTC: negative control. (B) Representative gel image used for the zygosity analysis of ACC1-specific insertions based on full-length PCR. The four sites on the left were homozygous L1Hs insertions and the others were heterozygous L1Hs insertions. (C) The distributions of signal counts (reads with unique start positions) per ACC1-specific insertion closely followed Poisson distributions (chi-squared goodness-of-fit tests). (D) Representative ACC1-specific insertion (ACC1_132 at chr21:29069173) in 1%, 0.1%, and 0.01% spike-in libraries. Read coverage and supporting signal counts (unique start positions were indicated by black arrows) were positively correlated with the spike-in concentration. (E) The effectiveness of error filters. 64 ACC1-specific germline insertions in 1%, 0.1%, and 0.01% spike-in libraries were considered as "true positives"; all other signals were considered as "false positives", which might include both background noise and some true somatic insertions present in the blood gDNA.

https://doi.org/10.1371/journal.pgen.1008043.g002

only 3.40% (66), 6.90% (181), and 7.70% (183) of false positives remained after all filters (S3 Table). These results showed that HAT-seq performed in combination with our error filters could successfully remove most artifacts and identify very low-frequency somatic insertions in bulk DNA samples.

Profiling of somatic L1Hs insertions in brain and non-brain human tissues

Next, we applied HAT-seq to 20 bulk samples obtained from postmortem neuronal (PFC neurons) and non-neuronal tissues (heart, eye, or fibroblast) from five Rett syndrome patients and five neurologically normal age-, gender-, and race-matched controls (Table 2 and S4–S7 Tables). A total of 9,181 putative somatic L1Hs insertions were identified in these 20 HAT-seq libraries (S8 Table). A subset of 137 (1.49%) of these insertions were detected by reads with multiple start positions. Considering that the random fragmentation process in HAT-seq library preparation would result in only one start position shared by all reads generated from a single cell, these 137 insertions should be present in multiple cells in the bulk tissue, and thus classified as "clonal somatic insertions". Based on the performance evaluation of HAT-seq, the lower bound of the precision of overall somatic L1Hs insertions was 60.14% (see details in Methods). To demonstrate the validity of these identified somatic insertions *in silico*, we investigated whether they had the hallmark features of TPRT-mediated retrotransposition (see details in Methods). By exploiting the sequence information of L1 integration junctions, we found that such somatic insertions were significantly enriched in genomic regions containing L1 endonuclease cleavage motifs (L1 EN motifs) ($p < 2.2\times10^{-16}$, Wilcoxon rank–sum test; Fig 3A; S4 Fig; S9 Table). Moreover, our identified somatic insertions shared the 25-bp peak of poly-A tail length with the reference L1Hs insertions (Fig 3B and S9 Table), where some of the somatic insertions with shorter tails might be explained by non-TPRT mechanism [31]. These features of somatic L1Hs insertions helped to elucidate the specificity of HAT-seq method.

Owing to the rarity of each somatic insertion in the cell population and to the sensitivity limits of various analytical methods, experimental validation of somatic insertions using unamplified bulk DNA, in particular when one of the primers is complementary to numerous homologous sequences in the human genome is very challenging (S4 Appendix). In theory, if a somatic insertion was unique to a single cell, it would be impossible to be detected in any replicated gDNA extracted from the same tissue. To circumvent this, we performed single-copy cloning by adapting a modified version of digital nested 3' PCR [10] that focused exclusively on clonal somatic insertions with three or more supporting signals, whose mosaicism (percentage of cells) were at least 0.1% based on our experimental design of HAT-seq library (Fig 3C). Five out of eight (62.5%) such clonal insertion sites were confirmed via 3' nested PCR and Sanger sequencing of cloned amplification products (Fig 3D–3H and S10 Table). Four of these clonal somatic insertions were located in introns of *TGM6*, *CNTN4*, *DIP2C*, and *DGKB*; three were sense-oriented to transcripts.

To our knowledge, no somatic insertions in non-brain tissues of healthy individuals has been reported [32]. We identified and experimentally validated a heart-specific somatic L1Hs insertion from a healthy individual UMB#1571 (Fig 3D). Leveraging both the 3' and 5' junctions of somatic L1Hs insertions enable us to characterize the terminal site duplications (TSDs) and L1 endonuclease cleavage site of insertion. Because most of somatic L1Hs insertions were 5' truncated with varied lengths, we screened and selected 22 high-quality step-wise primers covering the full-length L1Hs elements to capture their 5' junction (Fig 3C; S11 Table; S4 Appendix). Using 5' junction nested PCR, we successfully re-captured and Sanger sequenced the 5' junction of the heart-specific L1Hs insertion in the healthy individual (UMB#1571) (Fig 4A and S11 Table). We confirmed this insertion was a full-length somatic

Table 2. Overview of postmortem human tissues.

UMB ID[a]	Category	Mutation	Age	PMI	Brain tissue	Non-brain tissue	Matched ID	Number of somatic insertions		Rate of somatic insertions per cell		Known reference insertions (KRs)	Known non-reference insertions (KNRs)	Unknown insertions (UNKs)
								Cortical neuron	Non-brain	Cortical neuron	Non-brain			
4882	Rett syndrome	c.763C>T (p.R255X)	17 years 310 days	18 hrs	PFC	Heart	4591	855	364	1.47	1.01	819	194	10
1815	Rett syndrome	IVS3-2 A>G	18 years 130 days	5 hrs	PFC	Eye	1571	589	306	1.30	0.68	806	189	11
4852	Rett syndrome	c.451G>T (p.D151Y)	19 years 280 days	13 hrs	PFC	Eye	1347	380	257	0.89	0.54	824	171	8
4516	Rett syndrome	c.763C>T (p.R255X)[b]	20 years 356 days	9 hrs	PFC	Fibroblast	1846	759	411	1.82	0.69	814	195	7
1420	Rett syndrome	No pathogenic mutations	21 years 22 days	18 hrs	PFC	Heart	1455	708	216	1.31	0.37	824	182	8
4591	Healthy control	NA	16 years 223 days	14 hrs	PFC	Heart	4882	861	190	1.58	0.34	816	187	12
1571	Healthy control	NA	18 years 138 days	8 hrs	PFC	Heart	1815	384	221	0.63	0.57	813	175	11
1347	Healthy control	NA	19 years 76 days	16 hrs	PFC	Heart	4852	553	260	1.08	0.72	806	179	12
1846	Healthy control	NA	20 years 221 days	9 hrs	PFC	Heart	4516	744	296	1.66	0.68	817	175	8
1455	Healthy control	NA	25 years 149 days	7 hrs	PFC	Heart	1420	628	203	1.16	0.41	802	183	11

[a] The gender and race for all individuals were female and Caucasian.

[b] Genetic variant identified by custom AmpliSeq capture panel (S4 Table).

https://doi.org/10.1371/journal.pgen.1008043.t002

L1Hs insertion with 14 bp TSD and a cleavage site at 5'–TT/AAAG–3', similar to the consensus L1 EN motif 5'–TT/AAAA–3' (Fig 4B–4D). Notably, we also validated this 5' junction by combining full-length PCR with 5' junction PCR (Fig 4E; see details in Methods).

In addition, we verified one fibroblast- and another heart-specific L1Hs insertion in two patients with Rett syndrome (Fig 3E and 3F). The heart-specific L1Hs insertion in the Rett patient (UMB#1420) was further resolved to be a highly 5' truncated L1Hs insertion (~800 bp) with 9 bp TSD and a cleavage site at 5'–TT/TAAA–3' (S5 Fig and S11 Table). The poly-A tails

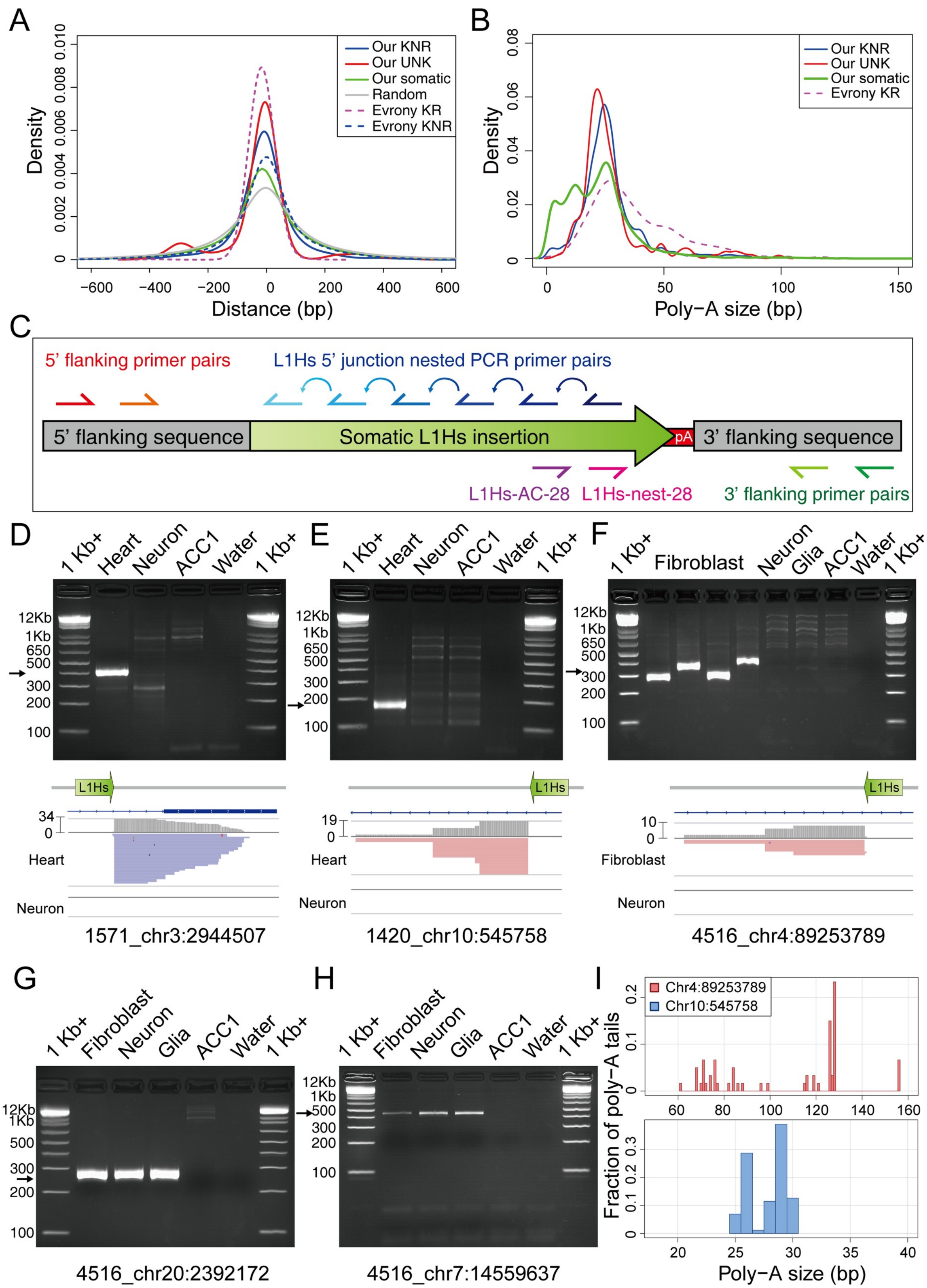
A
Density
0
0.002
0.004
0.006
0.008
0.010
−600
−400
−200
0
200
400
600
Distance (bp)
Our KNR
Our UNK
Our somatic
Random
Evrony KR
Evrony KNR
B
Density
0
0.02
0.04
0.06
0.08
0
50
100
150
Poly−A size (bp)
Our KNR
Our UNK
Our somatic
Evrony KR
C
5' flanking primer pairs
L1Hs 5' junction nested PCR primer pairs
5' flanking sequence
Somatic L1Hs insertion
pA
3' flanking sequence
L1Hs-AC-28
L1Hs-nest-28
3' flanking primer pairs
D
1 Kb+
Heart
Neuron
ACC1
Water
1 Kb+
12Kb
1Kb
650
500
300
200
100
L1Hs
34
0
Heart
Neuron
1571_chr3:2944507
E
1 Kb+
Heart
Neuron
ACC1
Water
1 Kb+
12Kb
1Kb
650
500
300
200
100
L1Hs
19
0
Heart
Neuron
1420_chr10:545758
F
1 Kb+
Fibroblast
Neuron
Glia
ACC1
Water
1 Kb+
12Kb
1Kb
650
500
300
200
100
L1Hs
10
0
Fibroblast
Neuron
4516_chr4:89253789
G
1 Kb+
Fibroblast
Neuron
Glia
ACC1
Water
1 Kb+
12Kb
1Kb
500
300
200
100
4516_chr20:2392172
H
1 Kb+
Fibroblast
Neuron
Glia
ACC1
Water
1 Kb+
12Kb
1Kb
500
300
200
100
4516_chr7:14559637
I
Chr4:89253789
Chr10:545758
Fraction of poly−A tails
0
0.1
0.2
60
80
100
120
140
160
0
0.1
0.2
0.3
20
25
30
35
40
Poly−A size (bp)

Fig 3. Profiling of somatic L1Hs insertions in multiple human tissues. (A) The density distributions of L1 EN motifs around L1Hs integration sites. L1 EN motifs included seven specific motifs (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, TCAAAA). "Evrony KR" and "Evrony KNR" are germline L1Hs insertions identified in Evrony *et al.* 2012. (B) The density distributions of poly-A tail length for each category of L1Hs insertion. (C) The PCR validation scheme and locations of primers used. (D)–(H) Representative gel images of 3' nested PCR validation for putative clonal somatic insertions. The Integrative Genomics Viewer screenshots for (D)–(F) showed the coverage track (gray) and the alignment track (blue for read strand [−]; red for read strand [+]) from HAT-seq data. Black arrows indicated bands with target size. 1Kb +: 1 Kb Plus DNA ladder. (I) Polymorphic poly-A tail sizes of clonal somatic insertions measured by capillary electrophoresis. Top: fibroblast-specific somatic L1Hs insertion at chr4:89253789 from Rett patient UMB#4516. Bottom: heart-specific somatic L1Hs insertion at chr10:545758 from Rett patient UMB#1420.

https://doi.org/10.1371/journal.pgen.1008043.g003

of these two clonal somatic insertions were experimentally measured to be polymorphic, indicating that they may involve multiple mutations after the original somatic retrotransposition events (Fig 3I and S10 Table). As previously reported [10, 33], poly-A tail was shown to be a highly mutable sequence element and might undergo secondary mutations in descendant cells. Furthermore, we confirmed two additional somatic L1Hs insertions from Rett patient UMB#4516 were present in PFC neurons, PFC glia, and fibroblasts (Fig 3G and 3H and S6 Fig), suggesting that they might retrotranspose during early embryonic development. Notably, the intronic somatic insertion (chr20:2392172) in *TGM6* was a full-length L1Hs insertion with 15 bp TSD and a cleavage site at 5'–AT/AAAA–3' (S7 Fig and S11 Table). We further quantified the allele fractions of this insertion using custom droplet digital PCR (ddPCR) assay and found that 6.34% of fibroblasts and 2.87% of PFC neurons contained this L1Hs insertion (S8A–S8E Fig and S10 Table). Our observations demonstrated that endogenous L1Hs could retrotranspose in various types of non-brain tissues during human development.

Abnormal L1Hs mobilization in patients with Rett syndrome

Our HAT-seq bulk sequencing data enabled us to perform statistical analysis of the exonic and intronic patterns of somatic L1Hs insertions in samples from Rett patients and matched healthy controls. We found 180 somatic insertions that were integrated into exonic regions: 9 of which were located in 5' UTR, 102 of which were located in coding regions, and 69 of which were located in 3' UTR (S12 Table). While no significant difference was observed in introns (odds ratio [OR] = 0.97, p = 0.44, Fisher's exact test), somatic insertions were significantly depleted in exons (OR = 0.59, p = 6.6×10^{-4}, Fisher's exact test) of Rett patients compared with matched healthy controls (Fig 5A and S13 Table). Previous studies have shown that dysregulation of long genes (> 100 kb) was linked to neurological disorders, including Rett syndrome [34] and autism spectrum disorder [35]. We used our HAT-seq data to investigate somatic insertional bias in both long (> 100 kb) and short genes (< 100 kb) of Rett patients. As a result, we found significant depletion of somatic insertions in exons of long genes (OR = 0.27, p = 5.2×10^{-5}, Fisher's exact test) but not short genes (OR = 0.76, p = 0.12, Fisher's exact test; Fig 5B and S13 Table). Our speculation was that if an L1Hs inserted into the exonic regions, especially in important genes, of the *MECP2* mutated cell, the cell would have a higher risk of death and subsequently be cleared up; therefore, the observed exonic depletion of L1 insertions in Rett patients might be resulted from the negative selection acting on those "lethal" exonic insertions.

In contrast to germline insertions, the impact of somatic insertions depends not only on their genomic location, but also the number of cells carrying that insertion, highlighting the importance of clonal somatic insertions. We found that in cortical neurons of Rett patients, clonal somatic insertions were enriched in introns (OR = 1.85, p = 0.029, Fisher's exact test; Fig 5C and S13 Table); these clonal intronic insertions were significantly enriched in the sense orientation to the transcripts (OR = 3.3, p = 0.0067, Fisher's exact test; Fig 5D and S13 Table). The presence of L1 insertion in the sense orientation has been reported to interfere with

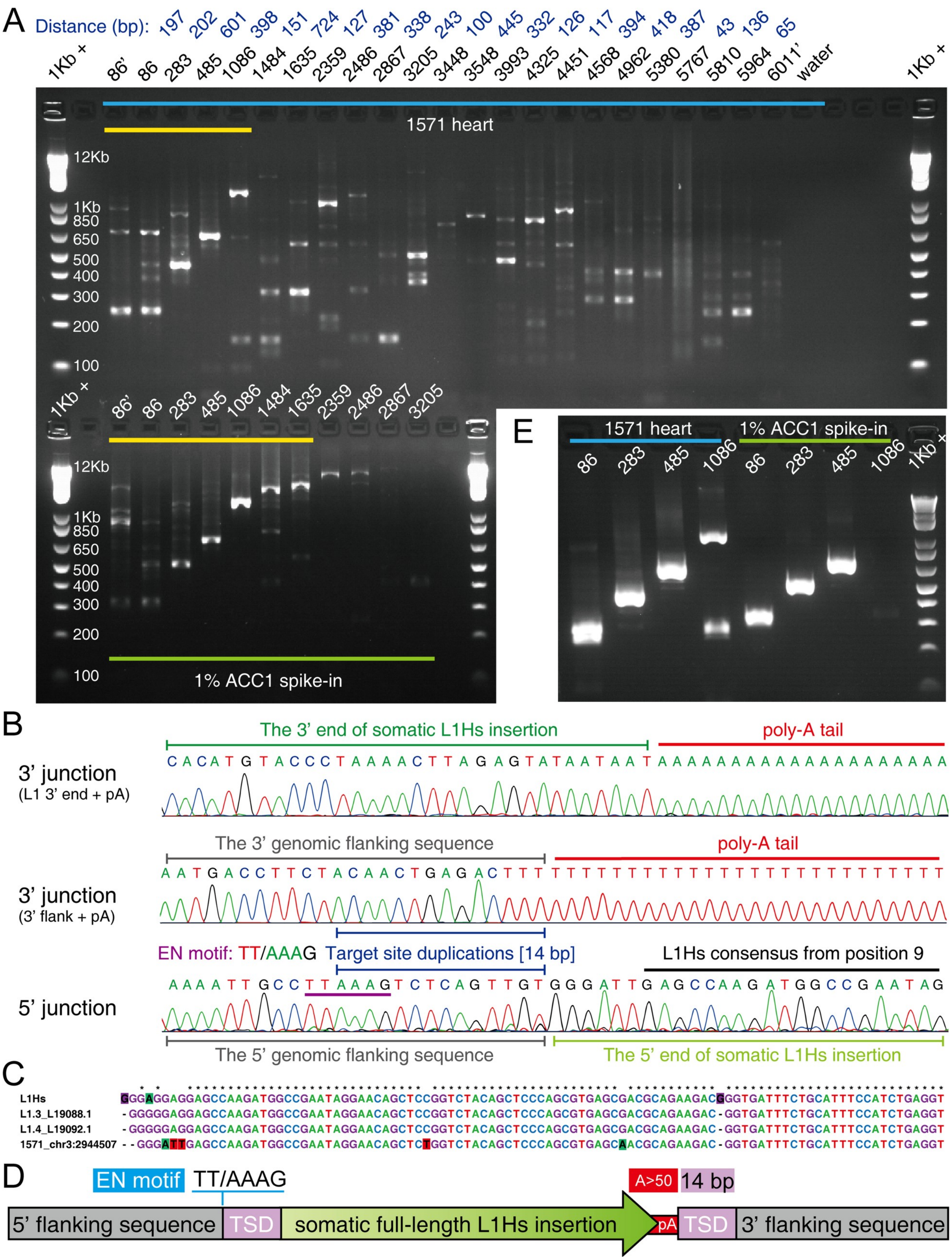
A
Distance (bp): 197 202 601 398 151 724 127 381 338 243 100 445 332 126 117 394 418 387 43 136 65
1Kb + 86' 86 283 485 1086 1484 1635 2359 2486 2867 3205 3448 3548 3993 4325 4451 4568 4962 5380 5767 5810 5964 6011' water 1Kb +
1571 heart
12Kb
1Kb
850
650
500
400
300
200
100
1Kb + 86' 86 283 485 1086 1484 1635 2359 2486 2867 3205
1% ACC1 spike-in
E
1571 heart
1% ACC1 spike-in
86 283 485 1086 86 283 485 1086 1Kb +
B
The 3' end of somatic L1Hs insertion
poly-A tail
3' junction
(L1 3' end + pA)
CACATGTACCCTAAAACTTAGAGTATAATAATAAAAAAAAAAAAAAAAAAAAA
The 3' genomic flanking sequence
poly-A tail
3' junction
(3' flank + pA)
AATGACCTTCTACAACTGAGACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
EN motif: TT/AAAG
Target site duplications [14 bp]
L1Hs consensus from position 9
5' junction
AAAATTGCCTTAAAGTCTCAGTTGTGGGATTGAGCCAAGATGGCCGAATAG
The 5' genomic flanking sequence
The 5' end of somatic L1Hs insertion
C
L1Hs GGGAGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACGCAGAAGACGGGTGATTTCTGCATTTCCATCTGAGGT
L1.3_L19088.1 -GGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACGCAGAAGAC-GGTGATTTCTGCATTTCCATCTGAGGT
L1.4_L19092.1 -GGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACGCAGAAGAC-GGTGATTTCTGCATTTCCATCTGAGGT
1571_chr3:2944507 --GGGATTGAGCCAAGATGGCCGAATAGGAACAGCTCTGGTCTACAGCTCCCAGCGTGAGCAACGCAGAAGAC-GGTGATTTCTGCATTTCCATCTGAGGT
D
EN motif TT/AAAG
A>50 14 bp
5' flanking sequence
TSD
somatic full-length L1Hs insertion
pA
TSD
3' flanking sequence

Fig 4. A full-length heart-specific L1Hs insertion (1571_chr3:2944507) in a healthy individual. (A) The agarose gel image of 5' junction nested PCR validation for the heart-specific L1Hs insertion in the healthy individual (UMB#1571; upper panel). The locations of primers used in 5' junction PCR assays were labeled on the top of each lane, where primers with the prime symbol denoted semi-nested PCR assays. The distances between each two adjacent 5' step-wise primers were labeled on the top (dark blue). The lower panel represented a heterozygous, full-length L1Hs insertion (ACC1_16; S11 Table and S4 Appendix) in 1% ACC1 spike-in gDNA as the positive control. The yellow line highlighted the expected stair-step bands in 5' junction PCR. 1Kb +: 1 Kb Plus DNA ladder. (B) The Sanger sequencing chromatograms of the 3' and 5' junctions of the somatic insertion 1571_chr3:2944507. The L1 EN motif and TSD were indicated by purple and blue lines. (C) Multiple sequence alignment of the 5' end between the identified somatic insertion and three L1Hs consensus sequences (L1Hs Repbase consensus and two hot L1s in human [L1.3 and L1.4]). (D) The schematic structure of 1571_chr3:2944507. (E) The agarose gel image of "full-length PCR + 5' junction PCR" assays for 1571_chr3:2944507 and ACC1_16 positive control.

https://doi.org/10.1371/journal.pgen.1008043.g004

transcriptional elongation of co-localized genes [36]. Considering that clonal insertions are more likely to have occurred at an early stage of development and thus affect a relatively large

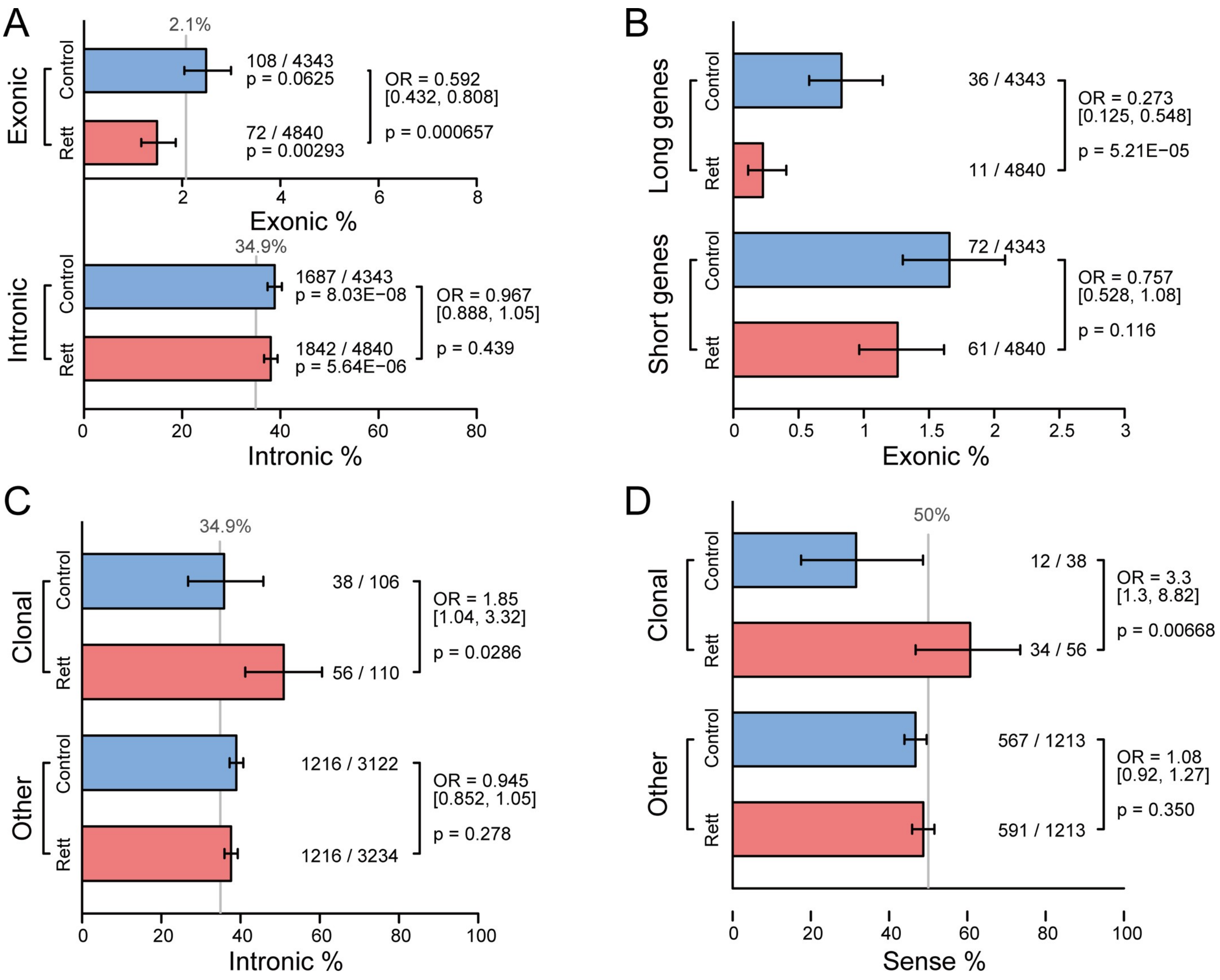

Fig 5. Abnormal L1Hs mobilization in patients with Rett syndrome. (A) Percentages of somatic L1Hs insertions in exons and introns. (B) Percentages of somatic L1Hs insertions in exons of long (> 100 kb) and short genes (< 100 kb). (C) Percentages of clonal somatic L1Hs insertions in introns. (D) Percentages of sense-oriented clonal somatic L1Hs insertions. The gray lines in (A) and (C) denoted the expected proportion determined by the exact base-pair count of that specific region relative to the human genome. The gray line in (D) represented the expected proportion if the insertions occurred randomly in both directions. Error bars in (A)–(D) indicated the 95% confidence intervals.

https://doi.org/10.1371/journal.pgen.1008043.g005 proportion of cells, these distinct insertion pattern in cortical neurons of Rett patients might indicate potential transcriptional burden on the nervous system.

Genomic patterns of somatic and germline L1Hs insertions

The design of HAT-seq method allowed for unbiased enrichment of both somatic and germline L1Hs insertions from each of bulk DNA samples. As germline insertion had constant genomic copy number in all tissues from the same donor, we used germline insertion as endogenous control to measure the relative copy number of genome-wide somatic insertions in the brain and non-brain tissues. We quantified the relative somatic L1Hs content by calculating the L1Hs-derived read count ratio of somatic to germline insertions using HAT-seq data of each sample (S14 Table; see details in Methods). Among all Rett patients and their matched controls, we observed a significant increase in the copy number of somatic L1Hs insertions in PFC neurons relative to matched non-brain tissues (heart, eye, or fibroblast) from the same donor ($n = 10$, $p = 2.7\times10^{-4}$, paired *t*-test; Fig 6A and S8F–S8G Fig). We also estimated the occurrence rate of somatic L1Hs insertions based on the germline insertion copy number of each individual (Fig 6B). This produced an average of 1.29 [95% CI: 1.03–1.55] somatic insertions per PFC neuron versus 0.60 [95% CI: 0.46–0.74] insertions per non-brain cell (S14 Table). Our observation of higher somatic L1Hs rate in PFC neurons from healthy individuals argued for the active retrotransposition of L1Hs in the human brain [9]. One significant advantage of HAT-seq was the ability to distinguish signals of somatic insertions from the overwhelming copies of germline L1Hs insertions in the genome (see details in Methods). Inconsistent with the previous qPCR result [8], when comparing the group of Rett patients with matched healthy controls, we only observed a slight but not significant increase of somatic L1Hs insertion rate in the Rett group, with 1.36 [min: 0.89; max: 1.82] versus 1.22 [min: 0.63; max: 1.66] per PFC neuron and 0.66 [min: 0.37; max: 1.01] versus 0.54 [min: 0.34; max: 0.72] per non-brain cell (Fig 6C and S14 Table).

We next characterized the genome-wide germline L1Hs insertions. HAT-seq yielded greater than 320-fold enrichment for KR, KNR, and UNK L1Hs insertions (S15 Table). On average, 814 KRs, 183 KNRs, and 10 UNKs were identified in each bulk sample (Table 2, S5–S7 Table). Hierarchical clustering based on L1Hs profiles correctly paired all neuronal samples with the non-neuronal tissue samples of the same individual (Fig 6D). To experimentally validate the HAT-seq predicted germline insertions, we performed 3' PCR validation on a random subset of polymorphic insertions from among the ten individuals, including 8 sites out of 160 polymorphic KRs, 20 sites out of 451 KNRs, and 2 sites out of 48 UNKs (S7 and S16 Tables). As a result, all of the assayed sites were detected in 3' PCR, with 98.4% (120/122) and 100% (168/168) sensitivity and specificity, respectively (S16 Table and S5 Appendix). These results support that HAT-seq can reliably detect germline L1Hs insertions with high sensitivity and specificity.

Previous studies have shown that intronic germline L1Hs insertions are sense-depleted [12, 25, 37]. As expected, the germline insertions identified in this study were significantly sense-depleted to the transcripts (633/1,544 [41%], $p = 1.6\times10^{-12}$, binomial test; Fig 6E and S13 Table). It is important to ask the question: whether such orientation bias for germline insertions is resulted from natural selection or insertional preference? To address this, we chose somatic L1Hs insertions as internal reference to control confounding factors. We compared the orientation bias between germline and somatic L1Hs insertions in transcripts and found that germline insertions were significantly sense-depleted than somatic insertions (OR = 0.79, $p = 7.9\times10^{-4}$, Fisher's exact test; Fig 6E and S13 Table). Because somatic L1Hs insertions only affected a small proportion of cells and thus they should undergo weaker selective pressure

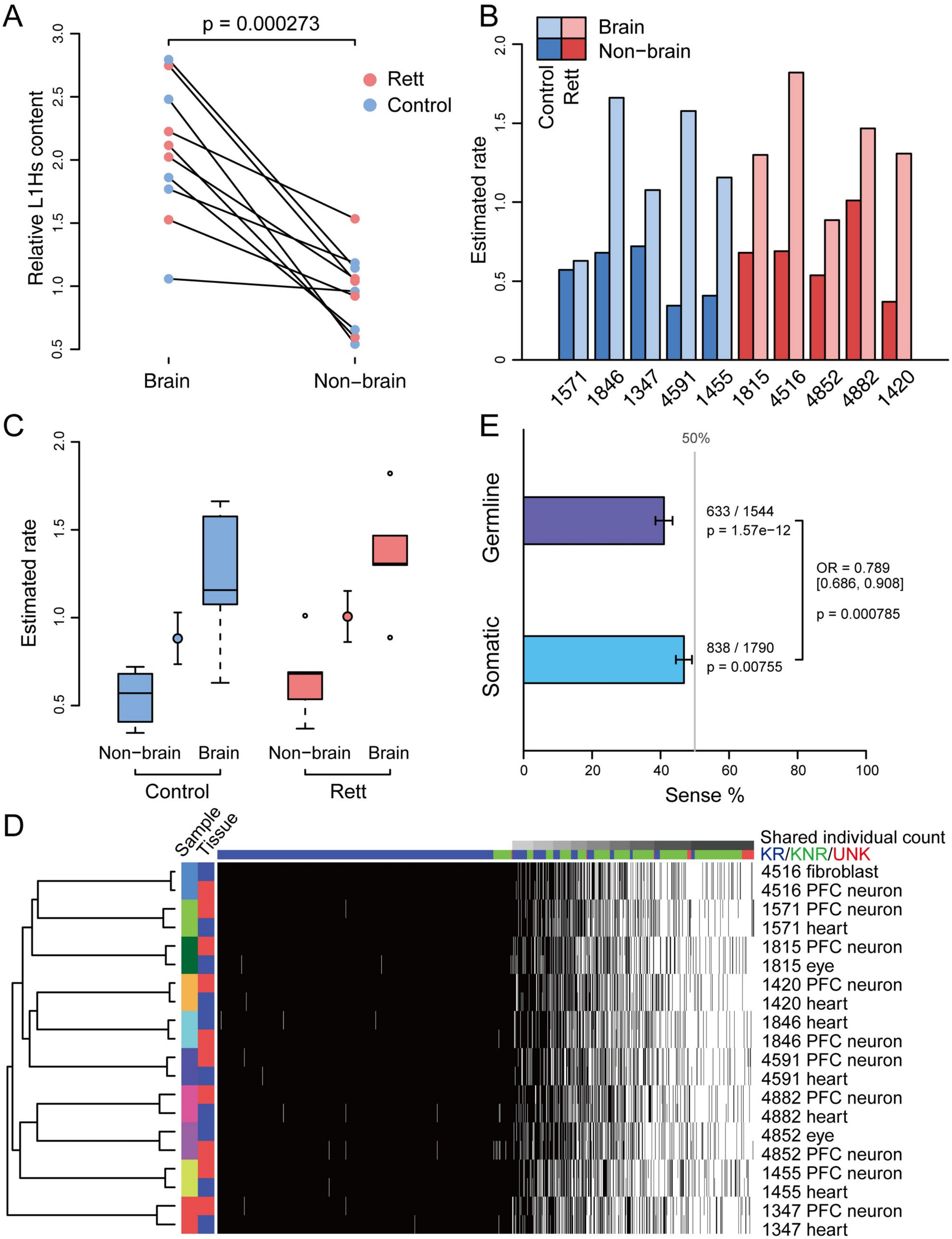

Fig 6. Genome-wide patterns of somatic and germline L1Hs insertions. (A) Relative somatic L1Hs content in PFC neurons and non-brain tissue from the same donor. The read count ratio of somatic insertions to germline KNR was calculated and then normalized relative to the average value

of non-brain samples. The linked dots represented pairs of brain and non-brain samples obtained from the same individual. (B) Histogram of estimated rate of somatic L1Hs insertions in each of tissue samples from the same donor based on the germline KNR copy number of each individual. (C) Estimated rate of somatic L1Hs insertions for different tissue types and cohorts. A circle with error bar denotes the estimate and the standard error of the mean (S.E.M) of all brain and non-brain samples. (D) Hierarchical clustering of all samples sequenced in this study. Each row represented a sample, and each column represents an L1Hs germline insertion. Black and white squares indicated the presence or absence of insertion, respectively. Column annotations showed categories for known reference (KR; blue), known non-reference (KNR; green), and unknown (UNK; red) insertions. (E) Percentages of sense-oriented germline and somatic L1Hs insertions in transcripts. The gray line represented the expected proportion if the insertions occurred randomly in both directions. Error bars indicated the 95% confidence intervals.

https://doi.org/10.1371/journal.pgen.1008043.g006

than germline insertions, our results suggested that natural selection may play a major role in shaping the sense-depleted distribution of germline L1Hs insertions.

Discussion

Here, we present HAT-seq, a bulk DNA sequencing method to profile genome-wide L1Hs insertions from physiologically normal and pathological human tissues. We demonstrated that, in addition to neuronal cells [7, 10–13], L1Hs also retrotransposed in a variety of non-brain tissues and cell types during normal development and contributed to the inter-cellular diversity of the human genome. Using high-throughput sequencing-based quantitative analysis, we found that somatic insertions occurred at a higher rate in brain than in non-brain tissues, consistent with previous studies [9]. Previous qPCR and single-cell genomic studies have resulted in conflicting estimates of the frequency of somatic insertions in neurons: ~80 L1 insertions per neuron [9], $<$ 0.04–0.6 L1 insertions per neuron [11], 13.7 L1 insertions per neuron [12], or ~0.58–1 somatic L1-associated variants per neuron [7]. Differential estimates might result from differences in WGA and signal enrichment methods. Using a bulk DNA sequencing approach, we estimated the rate of somatic insertions to be 0.63–1.66 L1Hs insertions per PFC neuron in healthy individuals (Fig 6B and S14 Table).

Clonally distributed insertions are prevalent in normal brain [10]. Increasing evidence suggests that neuronal L1s retrotransposition contributes to the susceptibility to and pathophysiology of neurological disorders, including Rett syndrome [8], schizophrenia [14] and Alzheimer's disease [38]. We observed that, in PFC neurons of Rett patients, clonal somatic insertions were enriched in introns, and these clonal intronic insertions were significantly enriched in the sense orientation (Fig 5C and 5D). In particular, in Rett patient UMB#4516, we found a full-length, sense-orientated, intronic somatic insertion (chr20:2392172) in *TGM6* (S7 Fig and S11 Table), a gene associated with central nervous system development and motor function [39], which could potentially dysregulate gene expression [36]. We found that 6.34% of fibroblasts and 2.87% of PFC neurons contained this insertion (S8A–S8E Fig and S10 Table), suggesting that it might occur in the 16-cell or 32-cell stages during morula stage. Mutations in *TGM6* are associated with spinocerebellar ataxia type 35, one of a group of genetic disorders characterized by poor coordination of hands, gait, speech, and eye movements as well as frequent atrophy of the cerebellum [40–42]. According to the clinical records, UMB#4516 had slight cerebral atrophy and cerebellar degeneration, could not hold things in her hands, and her speech development ceased at 16 months of age; these phenotypes were absent in the other four patients with Rett syndrome. Taken together, our data indicated that this clonal L1Hs insertion of *TGM6* might be correlated with the distinct clinical phenotype of UMB#4516.

Previous studies have provided evidence for significant selection against older L1 elements that are non-polymorphic [25, 43]. To characterize the insertion pattern of L1 with minimal influence from selective pressure, experimental methods were developed for recovery of novel L1 insertions in cultured cells [44, 45]. Using HAT-seq method, we were able to distinguish somatic L1Hs insertions from germline L1Hs insertions within the same individual. To determine whether the sense-depleted germline insertion was resulted from natural selection or insertional preference, we used somatic insertion as internal reference to control confounding factors such as intrinsic insertion preference and compared germline with somatic insertions. Our results suggested that natural selection shaped a sense-depleted distribution of germline L1Hs insertions in the human genome.

Several PCR-based bulk sequencing methods, such as ATLAS [24], L1-seq [25], TIP-seq [26], bulk SLAV-seq [7], and ATLAS-seq [27], have been developed to identify germline L1Hs insertions. Furthermore, L1-seq and TIP-seq have been successfully used in the identification of somatic insertions in tumors [26, 46–49]. Due to clonal expansion during tumorigenesis, such insertions could affect numerous cells in tumors. To our knowledge, HAT-seq is the first PCR-based bulk sequencing method to identify rare somatic insertions in a subset of cells—even unique cells—in non-tumor tissues. HAT-seq provides not only the genomic positions of somatic insertions but also the allele fraction of each insertion, which is informative for inferring the timing when the insertion has occurred. The sensitivities of HAT-seq for low-frequency somatic L1Hs insertions were relatively low (~30% for insertions present in $< 1\%$ fraction of cells). One possible explanation was that some signals of insertion were lost during library construction and NGS sequencing, e.g. sonic fragmentation, clean-ups, size selection, and loading library to sequencer. Single-cell whole genome and targeted sequencing approaches have been used to identify both TPRT-mediated and endonuclease-independent insertions [7, 10–12], where the signal of somatic insertions can be as high as germline heterozygous insertions in single-cell level. However, such single-cell approaches cannot achieve increased sensitivity without cost [22]. For example, to detect a given insertion with 0.1% mosaicism, more than 1,000 single cells may need to be amplified and sequenced. Therefore, compared with single-cell approaches, HAT-seq was eligible to identify a large number of somatic L1Hs insertions in a more cost-effective way.

Based on our experimental design, assembling overlapped read pairs into contigs can provide sequence information fully spanning the L1Hs integration sites, enabling downstream false-positive filtering based on both sequence features and read-count. However, a portion of read pairs were unable to be merged into contigs because of the inaccurate size-selection during library construction. Applying the same filtering strategy, we re-analyzed these unassembled read pairs and revealed 11 clonal insertion candidates. Further PCR experiments only validate one of these candidates (9%, S10 Table). Because the key filter “chimera within poly-A tail” was not applicable for unassembled read pairs, our sequence analysis suggested that chimeric molecules bridging within the poly-A tail was the major source of false-positives for unassembled data (see details in Methods). As shown in the statistics of positive control libraries (S2 Table) and experimental validation, the unassembled data could provide additional signals of somatic L1Hs insertions but require careful analysis and rigorous validation to address technical artifacts. Further gains in statistical power will be benefited from increased sample size and improved efficiency of HAT-seq.

Several unresolved technical challenges might constrain the total number of detectable L1Hs insertions by the current version of HAT-seq, including the identification of insertions in repetitive regions with low mappability (such as pre-existing L1 germline insertions) and 3’ truncated insertions. With rapid innovations in sequencing technology, higher throughput and longer read length will markedly improve the performance of HAT-seq. Future studies that profile all active retrotransposons (i.e., L1Hs, Alu, and SVA) in a variety of cell types, tissues, and developmental stages will shed new light on the dynamics of somatic retrotransposition under host regulation and help to uncover their roles in human disease.

Methods

Ethics statement

Postmortem samples of prefrontal cortex and non-brain tissues were obtained from five patients with Rett syndrome (UMB#4882, UMB#1815, UMB#4852, UMB#4516, and UMB#1420) and five age-, gender-, and race-matched neurologically normal individuals (UMB#4591, UMB#1571, UMB#1347, UMB#1846, and UMB#1455) through the UMB Brain and Tissue Bank (University of Maryland, Baltimore, MD) (Table 2 and S4 Table); written informed consent was obtained by the UMB Brain and Tissue Bank and the Lieber Institute for Brain Development (Baltimore, MD). The peripheral blood samples of two unrelated individuals (ACC1 and ACC2) were collected with written informed consent by Peking University. This study was approved by the Institutional Review Board (IRB) at Peking University (IRB00001052-13025).

Isolation of single neuronal nuclei

Nuclei were isolated and labeled for FACS based on a previous study [14] with modifications. Fresh-frozen samples were thawed gradually from liquid nitrogen by transferring to a −80˚C freezer; the samples were then transferred to a −20˚C freezer 1 h later. All procedures were performed at 4˚C unless noted otherwise. First, 100 mg of tissue was minced into pieces, and transferred to 2 mL STKM buffer (250 mM sucrose, 50 mM Tris-HCl, pH 7.4, 25 mM KCl, 5 mM $MgCl_2$) with protein inhibitor (cOmplete, Mini, EDTA-free Protease Inhibitor Cocktail, Roche). The minced tissue was soaked overnight for 8 hours and homogenized in a Potter-Elvehjem glass homogenizer (886000–0019; Kontes).

To improve immunostaining and the purity of the isolated target, debris was removed by Percoll density gradient centrifugation. Brain homogenate was filtered through a 100-μm cell strainer and mixed with Percoll solution (P1644-100ML; Sigma) to a final concentration of 19%. A 5-mL ultracentrifuge tube (P/N 344057; Beckman) was layered with Percoll solutions in the following order: 0.4 mL of 12% Percoll, 3 mL of homogenate (19% Percoll), 0.8 mL of 26% Percoll, and 0.3 mL of 35% Percoll. The tube was then centrifuged in a SW 55 Ti rotor (Beckman Coulter) at 16,000 rpm (30,000 *g*) for 10 min. Large quantities of myelin and cellular debris generated during brain homogenate preparation were removed from the single-nuclei suspension, and the floating nuclei fraction was collected from the 35% layer (S1G–S1I Fig and S2D and S2E Fig).

Neuronal nuclei were purified using NeuN immunostaining. The nuclei fraction was blocked in 2.5% bovine serum albumin (BSA) in phosphate buffered saline (PBS) for 2 hours with 6 rpm end-to-end rotation; 20 μL of the sample served as an unstained control sample for flow cytometry. Blocked nuclei were labeled with 2μL/mL PE-conjugated anti-NeuN antibody (FCMAB317PE; Millipore), filtered through a 40-μm cell strainer and diluted with 1% BSA in PBS at 2 volumes of the sample. We did not stain the nuclear fraction with a fluorescent nuclear stain (e.g., propidium iodide [PI] or 4’,6-diamidino-2-phenylindole [DAPI]) for sorting because they bind to DNA and affect quantification analysis using Qubit 2.0 Fluorometer (Life Technologies). Single-nuclei suspension was sorted in 4-way purity mode at a flow rate less than 6,000 events per second using an 85-μm nozzle with a BD FACSAria II cell sorter. The collection tube was pre-coated with 1% BSA in PBS, and a small volume of 1% BSA in PBS was then pre-added to protect the nuclei from breaking down. Sorted $NeuN^+$ and $NeuN^-$ fractions were re-analyzed by flow cytometry to verify the purity; a small portion was stained with DAPI or PI to check the purity and integrity via fluorescence and differential interference contrast (DIC) microscopy. Nuclei were pelleted at the bottom of the collection tube after centrifugation with a swing rotor at 1,000 *g* at 4˚C for 20 min. gDNA was extracted using QIAamp DNA Mini Kit (QIAGEN) or QIAamp DNA Micro Kit (QIAGEN) according to the sorting statistic.

HAT-seq library construction and sequencing

First, 500 ng of gDNA was sonicated using Covaris S220 with the following settings: sample volume, 50 μL; water level, 12; temperature, 7˚C; peak incident power, 175 W; duty factor, 5%; cycles per burst, 200; and treatment time, 55 s. DNA fragments were end-repaired, dA-tailed, and adaptor-ligated using KAPA LTP Library Preparation Kit (KK8232; KAPA Biosystems). All oligonucleotides used in library preparation were synthesized by Invitrogen (Life Technologies) and are listed in S1 Table. Adaptor-ligated DNA (20 ng, ~3,000 cells) served as input for PCR-based target enrichment. The PCR protocol was: 12.5 μL KAPA2G Robust HotStart ReadyMix (2×) (KK5702; KAPA Biosystems), 1.25 μL P7_Ns_L1Hs (10 μM), and PCR-grade water added to final volume of 23.75 μL. Another primer, 1.25 μL P5_extension (10 μM), was added when linear amplification was finished. P7_Ns_L1Hs (10 μM) was an equimolar mixture of P7_N2_L1Hs, P7_N4_L1Hs, and P7_N6_L1Hs. The cycling programs were: 95˚C for 5 min; 5 cycles of 95˚C for 40 sec, 61˚C for 15 sec, and 72˚C for 15 sec; a pause at 12˚C to add the P5_extension primer; 11 cycles of 95˚C for 40 sec, 61˚C for 15 sec, and 72˚C for 15 sec; ending with 72˚C for 30 sec and held at 4˚C.

Post-PCR cleanup was performed with 1.05× Agencourt AMPure XP beads (Beckman Coulter, Inc.). Amplified products from the first PCR were eluted in 10 μL of Buffer EB (QIAGEN) and used as template in the second PCR to incorporate Illumina sequencing adapters with barcode. The PCR protocol was 12.5 μL KAPA2G Robust HotStart ReadyMix (2×), 1.25 μL P5_to_end (10 μM), 1.25 μL P7_extension_i7_index (10 μM), and PCR-grade water to final volume of 25 μL. The cycling program was: 95˚C for 5 min; 5 cycles of 95˚C for 40 sec, 60˚C for 15 sec, and 72˚C for 15 sec; ending with 72˚C for 30 sec and held at 4˚C.

To deal with “bubble products” from overamplification that could hinder accurate gel-based size selection, a step of “one-round PCR” was performed by adding equal volumes of KAPA2G Robust HotStart ReadyMix (2×), P5_to_end (10 μM), P7_extension_i7_index (10 μM), and PCR-grade water added to the PCR tube to a final volume of 50 μL. The cycling program was: 95˚C for 80 sec, 60˚C for 30 sec, 72˚C for 2 min, and held at 4˚C. Post-PCR cleanup was performed with 1.1× Agencourt AMPure beads. Each library was eluted with 30 μL of Buffer EB and size selected (320–550 bp) using Pippin Prep (Sage Science). After library quality control using Agilent 2100 Bioanalyzer with High Sensitivity DNA Kit (Agilent Technologies) and KAPA Library Quantification Kit Illumina platforms (KK4824, KAPA Biosystems), HAT-seq libraries were paired-end sequenced (2*150 bp) at Novogene, Inc.

Read alignment and peak calling

For 20 HAT-seq libraries constructed from postmortem human tissues, a total of 1,191,889,370 2*150 bp read pairs were generated, with an average of 59,594,469 read pairs per sample. Library details are shown in S5 Table. Schema of the HAT-seq data analysis pipeline is shown in S3B Fig.

Raw data were de-multiplexed, adaptor trimmed, and base trimmed with base quality $<$ 10. Next, we specifically extracted L1Hs-derived read pairs based on the Read 2 sequence information. Only those with the 3’ consensus sequence of P7_Ns_L1Hs “GGGAGATATACCTAA TGCTAGATGACAC” were retained and trimmed, as they had the correct HAT-seq library structure. Read 2 sequences with 95% identity to the 3’ end of the L1Hs consensus sequence were retained. For KR insertions, the retained Read 2 and their paired Read 1 were aligned to hg19 using BWA-MEM (version: 0.7.12-r1039, default parameters) and those uniquely mapped reads were used for peak calling. For non-reference insertions, we first merged L1Hs insertion-derived read pairs into contigs using PEAR (version: 0.9.6; parameters: -y 50M -j 4 -m 300 -n 70) and aligned contigs to hg19 with BWA-MEM (default parameters) which allowed for split-read mapping. All uniquely mapped contigs without any clipping were ignored in the downstream analysis as these contigs were deemed to be KR-derived reads. Only those clipped contigs and non-uniquely mapped contigs (contigs with mapQ $<$ 20 or unmapped contigs) were extracted for further computational analysis to call non-reference insertions. These contigs were poly-T (TTTTTTTT) trimmed, leaving 3' genomic flanking sequence of L1Hs insertions for STAR mapping to hg19 (version2.4.2a; parameters:—outFilterMatchNminOverLread 0.3—outFilterScoreMinOverLread 0.3—scoreGapNoncan -4—scoreGapGCAG -4—scoreGapATAC -4—alignIntronMax 500). For uniquely mapped reads, we marked PCR duplicates with SAMBLASTER (version: 0.1.22) and used them to call non-reference insertion peaks.

Peak calling was triggered where a genomic position with $>$1 depth was found. Adjacent peaks were merged with a maximum distance of 100 bp. We intersected peaks of each sample with the L1Hs insertions collected in RepeatMasker (database version: 20130422; http://www.repeatmasker.org) and annotated each overlapped peak with a total of 6 features: peak height (reads per million mapped reads [RPM]), peak width (genomic length with read depth $\geq$ 1), signal count (the number of unique start positions of reads aligning to the peak), depth of each signal (the number of PCR duplicates for each signal with unique start position), genomic information of overlapped L1 elements, and its overlapping width.

KR peak classification

We employed a read-count filter to distinguish true insertions from artifacts. Many false-positive KRs were supported by reads aligned to the 3' end of L1Hs with sufficient depth but without reads mapping to their 3' flanking sequence. Some false positives were supported by a few chimeric molecules with low depth. Putative KR peaks of each sample were assigned when they satisfied the following criteria: a) overlapped with annotated L1Hs regions in the human reference genome; b) RPM $>$ 40; c) overlapping width $>$ 200 bp; and d) not in chromosome Y (chrY).

KNR peak classification

Peaks were called and merged as described above, with the exception that they were performed separately for reads aligning to the plus and minus strands of the reference genome since the 3' flanking sequence preserved the insertional orientation information. We filtered out peaks that overlapped with reference L1Hs and L1 subfamilies (L1PA2, L1PA3, and L1PA4). We intersected the remaining peaks with the meta retrotransposon insertion polymorphisms (MRIP) list from euL1db [30], and assigned overlapped peaks as putative KNRs when they satisfied the following criteria: a) signal counts $\geq$ 30; b) RPM $\geq$ 100; c) at least 3 signals with a "depth of each signal" $\geq$ 5; and d) not in chrY.

UNK peak classification

For the remaining non-reference insertions, we implemented a series of empirical error filters to deliberately remove several types of false positives. First, we rejected reads with risk of misalignment, defined as when the BWA-MEM and BLAT alignments were inconsistent. Second, we rejected reads without an L1Hs diagnostic G motif (position 6012 relative to the L1Hs Repbase consensus). Third, we rejected reads at risk of being a chimeric molecule. We applied

BLAST to find the best alignments for retrotransposon and non-retrotransposon segments from hg19. Reads were removed as a putative chimera when the sequences of two segments overlapped > 10 bp with A% $\geq 50\%$ or overlapped 6–10 bp with A% $< 50\%$ [12]. Fourth, we rejected reads with risk of being derived from nearby reference L1Hs. We extracted 2 kb downstream of aligned non-retrotransposon segment from hg19 and aligned the full contig against this sequence by BLAT to exclude potential genomic rearrangement events [12]. To circumvent the interference of background noise, all contigs that passed these filters were extracted and remapped using STAR to acquire clean bam files for subsequent statistical analysis. Peaks were classified as putative UNKs when they satisfied the following criteria: a) signal counts ≥ 10; b) at least 3 signals with a "depth of each signal" ≥ 5; and c) not in chrY.

Somatic insertions classification

After filtering out all putative UNKs, most remaining non-reference insertions were supported with low read depth. To distinguish somatic insertions from artifacts, we regarded PCR duplicates as a marker of high-confidence somatic insertions. The rationale was that each L1Hs insertion in the template gDNA was amplified by 17 PCR cycles (11 + 5 + 1) and a portion of their duplicates should be sequenced. In contrast, technical artifacts induced by non-specific or chimeric PCR amplification were inevitable but were generated at a much lower rate. Therefore, we rejected putative peaks without PCR duplicates. As shown in S5 Table, we suggested to sequence at least 50M reads for each HAT-seq library. Finally, we rejected systematic error-prone sites shared by two or more individuals because the likelihood of recurrent somatic insertions in different individuals was presumed to be much smaller than the likelihood of systematic mapping or sequencing errors.

A subset of putative somatic insertions was classified as clonal somatic insertions, which were supported by two or more PCR duplicate signals with different unique start positions. If two unique start positions differed by a shift of 1 bp, we tolerated the difference and regarded them as the same signal, for this was likely due to low base quality at the beginning of Read 1 (3' flanking genomic sequence).

Peak joining across samples

Merged peak references were created for each of four insertion categories (KR, KNR, UNK, and somatic insertions). Peaks were merged with a maximum distance of 100bp in a strand-specific manner. Detailed information of germline and somatic insertions across all samples was provided in S5–S8 Tables. Hierarchical clustering of germline insertions across all samples was performed with heatmap3 package in R (https://www.r-project.org).

Identification of somatic insertion using unassembled paired-end reads

Due to the wide range of size-selection during library construction, a portion of unassembled read pairs in the HAT-seq data were filtered after contig merging step. We confirmed that additional signals of somatic L1Hs insertions exist in these unassembled data. As shown in S2 Table, higher sensitivity of HAT-seq will be achieved when including unassembled reads in HAT-seq analysis. These unassembled paired-end reads do not contain the sequence information of L1-genome junction and thus were not applicable to distinguish true signals (somatic L1Hs insertions) from noise (chimeric molecules) based on L1 integration site sequence features (hallmarks of TPRT mechanism). Applying the same criteria for "contig data" analysis except for the two inapplicable filters ("chimera within poly-A tail" and "local SV"), we identified 11 clonal somatic insertion candidates with three or more supporting signals, whose mosaicism (percentage of cells) were at least 0.1% based on our experimental design of HAT-seq library. Among the 11 putative clonal somatic insertions, only one event was confirmed via 3' nested PCR and Sanger sequencing (S10 Table [unassembled data]). This clonal somatic insertion (1571_chr3:2944507) has the maximum count of supporting signal and was supported by both "unassembled data" and "contig data" with 6 and 18 supporting signals (reads with different start positions), respectively. All false positives from unassembled data were overlapped with repeat elements, especially the 3' end of Alu subfamilies. Eight out of the ten false positives (80%) were supported by only three supporting signals. Notably, five false positives were identified from PFC neuron of UMB#4852. This library contained higher proportion of unassembled reads than other libraries (S5 Table [Uniquely mapped polyT trimmed read pairs using STAR]), indicating a higher level of background noise in this library.

Quantification of somatic L1Hs insertions

Previous studies reported relative quantification of L1Hs contents using TaqMan quantitative PCR (qPCR) [8, 9]. However, except for its limitation by using exogenous L1 plasmid to estimate L1 copy number [50, 51], the qPCR assay lacks specificity for active L1 elements [22]. On one hand, L1 reverse transcription occurring in cytoplasm would confound quantification [52]. On the other hand, the qPCR assay was unable to distinguish between somatic and germline L1Hs insertions, while L1Hs copy number variation among tissues was only contributed by active somatic insertions.

In HAT-seq library, genomic fragments containing L1Hs insertions served as templates and were amplified equally using the same PCR reaction conditions. A random subset of the library was subsequently sequenced and classified into somatic and germline insertion-derived reads. We quantified the relative copy number of somatic L1Hs in each tissue from the same donor by normalizing somatic L1Hs-derived read counts with germline L1Hs-derived read counts. As KNRs shared the same sequence features and non-reference insertion calling pipeline with somatic insertions, we further quantified the rate of somatic L1Hs insertions per cell based on the KNR copy number of each individual (S14 Table). Given that most KNRs are heterozygous, we regarded the KNR copy number of each individual as the KNR count per cell. To demonstrate the linear PCR amplification of both germline and somatic insertions during HAT-seq library construction, we calculated the estimated rate of 64 ACC1-specific spike-in insertions in the positive control libraries. Our observed rates were 0.59, 0.056, and 0.0144, approximated to the expected 0.64, 0.064, and 0.0064 ACC1-specific insertions per cell in 1%, 0.1% and 0.01% spike-in libraries, respectively (S14 Table).

Genomic annotation and statistical analysis

Using R and Bioconductor (https://www.bioconductor.org/) packages, we downloaded refGene annotations for the hg19 genome from UCSC Genome Browser and annotated the L1Hs insertions. If a gene produced multiple transcripts, we focused on the canonical transcript, the longest transcript among those with the longest coding sequence. We annotated the genomic coordination of each category of L1Hs insertions along with the human genome and applied a binomial test to compare the proportion of insertions located inside a specific region (e.g. introns or exons) with the expected proportion determined by the exact base-pair count of that specific region relative to the human genome. We also annotated the sense or antisense orientation of the peaks located in transcripts and applied a binomial test to compare the sense proportion with the expected 50% under null hypothesis. When comparing the intronic or exonic insertion proportion between Rett patients and control samples, we made a 2 × 2 contingency table and applied Fisher's exact test; we reported the p-value, the estimate and 95% CI of OR. Similarly, Fisher's exact tests were also applied to judge the difference in sense-oriented count between germline and somatic insertions. Statistical analysis on clonal and "unique" somatic insertions were performed using Fisher's exact test based on their count of insertion events. All annotations and statistical analyses were conducted by an automatic pipeline in R language to ensure reproducibility.

Distribution of distance to L1 EN motifs and poly-A tail length

Based on the strand of insertion, we retrieved 2-kb upstream and downstream genomic sequences of each of L1Hs integration sites and calculated their minimum distance from the integration sites to one of seven typical L1 EN motifs in TPRT-mediated retrotransposition (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, and TCAAAA) [53]. For random sampling control, we retrieved the upstream and downstream flanking sequences of 100,000 random positions in the genome. Sequences consisting of >20% low mappability nucleotides (mappability < 0.25) were removed (a total of 8,634 sequences). Furthermore, 67 sequences consisting of >90% N were removed. We applied Wilcoxon rank-sum test to compare the absolute values of the distances between each of insertion categories. For S4 Fig, we reduced the bin size to 10bp and illustrated the y-axis with a gap break.

In addition, we examined the poly-A tail length size of each L1Hs insertion. Due to our experimental design, the supporting contigs contained poly-T before the L1Hs sequence (reverse complementary). Therefore, we determined the anchor position of ATTAT on each contig using a greedy algorithm, and then calculated the poly-T length by searching backward to the upstream anchor position using a scoring algorithm (match [T] +1, mismatch −2, report poly-T with score ≥ 0). Starting from the anchor position at the end of L1Hs 3' UTR, our algorithm examined the poly-A tail length of L1Hs elements regardless of whether the retrotransposition event had 3' transductions or not. All annotations and statistical analyses were conducted using Perl 5 (https://www.perl.org) and R 3.1.0.

Positive control experiments

Because both polymorphic germline and somatic insertions belonging to L1Hs sharing the same sequence characteristics, known concentration spike-in of polymorphic germline insertions can be regarded as true somatic insertions with known allele frequency. First, we extracted DNA from the blood of two unrelated adults, ACC1 and ACC2. We identified 172 non-reference L1Hs insertions in ACC1 using its HAT-seq data and screened them in both ACC1 and ACC2 using 3' PCR. 64 polymorphic L1Hs insertions confirmed to be ACC1-specific (Fig 2A and S2 Appendix). The zygosity of ACC1-specific insertions was confirmed by full-length PCR with 49 (77%) heterozygous insertions, 9 (14%) homozygous insertions, and 6 (9%) zygosity-undetermined insertions (Fig 2B and S2 Appendix). As most ACC1-specific insertions were heterozygous, we used ACC1 gDNA as the spike-in, with 0.5 insertions per genome for each of ACC1-specific insertions.

A mixed-DNA series containing 1%, 0.1%, and 0.01% ACC1 gDNA were prepared using ACC1 and ACC2 gDNA. Using 20 ng (3,000 cells) adaptor-ligated gDNA as input, HAT-seq libraries were constructed, sequenced, and computationally analyzed in the same manner as the bulk sequencing HAT-seq libraries mentioned above, with the exception of ignoring the KNR insertion filtering. We labeled insertion sites as "detected" when they were supported by uniquely mapped reads, and "identified" when they were subsequently supported by reads that passed all stringent error filters.

64 ACC1-specific insertions were independent events. For each insertion, supporting signals could be counted based on different start positions. Because all the insertion-supporting reads originating from a single cell should have identical start position, the signal count of

each insertion indicated the number of cells carrying the insertion that were sampled from the library input. The distributions of supporting signal counts (reads with unique start positions) per ACC1-specific insertion should follow Poisson distribution. The parameter lambda for Poisson distribution was fitted using the maximum likelihood method, and chi-squared goodness-of-fit tests were performed (Fig 2C and S2 Table). For 1%, 0.1%, and 0.01% ACC1 spike-in libraries, each of 64 ACC1-specific insertions was diluted to 30, 3, and 0.3 copies. Theoretically, by Poisson statistics, there would be 64, 60.81, and 16.59 ACC1-specific insertions being sampled and subsequently being used as the input of HAT-seq libraries. According to the number of ACC1-specific insertions identified in 1%, 0.1%, and 0.01% libraries, the sensitivities for somatic L1Hs insertions were 76.6% (49/64), 28% (17/60.81), and 30.1% (5/16.59), respectively. As shown in S3 Table, the number of false positives were 66, 181, and 183 in 1%, 0.1%, and 0.01% spike-in libraries, respectively. The numbers of false positives were stable in 0.1% and 0.01% libraries, whereas the lower number of false positives in 1% library should result from the lower total throughput of the library (S5 Table). The percentage showed a 3.61-fold decrease (from 14.3% to 3.96%) after applying the "PCR duplicate" filter, compared to 2.28-fold decrease in 0.1% (from 17.23% to 7.55%) and 0.01% (from 19.09% to 8.37%) libraries, suggesting that more candidates without PCR duplicates were filtered in the 1% library due to lower sequencing throughput. In sum, our results suggested that the upper bound of the number of false positives should be 183 per library. Using the most stringent definition, the total number of false-positives in all 20 libraries was 3,660 out of 9,181 and thus the overall precision for our putative somatic L1Hs insertions was 60.14%.

3' PCR and full-length PCR for germline L1Hs insertions validation

To identify ACC1-specific insertions, the 3' PCR protocol comprised 10 ng template DNA, 10 μL 2× Taq PCR StarMix with Loading Dye (A012; GenStar), 1 μL site-specific 3' primer (10 μM), 1 μL L1Hs-AC-28 (10 μM), and PCR-grade water to 20 μL. The cycling program was: 94˚C for 2 min; 35 cycles of 94˚C for 30 sec, 60˚C for 30 sec, 72˚C for 30 sec; 72˚C for 5 min; and held at 4˚C. Validation primer sequences used for each candidate insertion can be found in S2 Table.

To determine the zygosity of ACC1-specific insertions, the full-length PCR protocol comprised 50 ng template DNA, 1 μL PrimeSTAR GXL DNA polymerase, 5 μL PrimeSTAR GXL buffer (5×), 2 μL dNTP mixture (2.5 mM each), 0.5 μL dimethyl sulfoxide (DMSO), 0.75 μL 5' primer (10 μM), 0.75 μL 3' primer (10 μM), and PCR-grade water added to a final volume of 25 μL. The cycling program was 98˚C for 3 min; 30 cycles of 98˚C for 15 sec, 58˚C for 20 sec, 68˚C for 2 min; 68˚C for 3 min; and held at 4˚C. Validation primer sequences used for each candidate insertion can be found in S2 Table.

For validation of polymorphic germline L1Hs insertions, the 3' PCR protocol comprised 10 ng template DNA, 10 μL KAPA2G Robust HotStart ReadyMix (2×), 1 μL site-specific 3' primer (10 μM), 1 μL L1Hs-AC-28 (10 μM), and PCR-grade water to a final volume of 20 μL. The cycling parameters were 94˚C for 3 min; 35 cycles of 94˚C for 15 s, 58˚C for 15 s, 72˚C for 15 s; 72˚C for 3 min; and held at 4˚C. Validation primer sequences used for each candidate insertion can be found in S16 Table.

All PCR products were run on 1.5% or 2% agarose gels, and images were analyzed using Image Lab software (Bio-Rad) to quantify the product sizes.

PCR assays for somatic L1Hs insertions validation

L1 3' digital nested PCR validation. To validate somatic insertions from bulk DNA, we adapted the 3' digital nested PCR (dnPCR) method [10] to validate somatic insertions from bulk DNA. Only clonal somatic insertions with three or more signals, whose mosaicism (percentage of cells) was at least 0.1% based on HAT-seq data, were considered for validation because these insertions should be present in the cell population with a low allele fraction; further, re-sampling from the bulk tissue gDNA was possible. Briefly, we assumed that the frequency of clonal somatic insertion was 1/2,000 cells. Then, each DNA sample was diluted to a target somatic insertion concentration of 0.3 copies/μL. The diluted DNA was run through two-round 3' nested PCR. The PCR protocols were as follows. Round 1 PCR (per reaction): 1 μL diluted DNA, 5 μL KAPA2G Robust HotStart ReadyMix (2×), 0.5 ×L site-specific 3' primer (10 ×M), 0.5 ×L L1Hs-AC-28 (10 ×M), and PCR-grade water to 10 ×L. The cycling program was: 94˚C for 3 min; 15 cycles of 94˚C for 15 sec, 58˚C for 15 sec, 72˚C for 15 sec; 72˚C for 2 min; and held at 4˚C. Round 2 PCR (per reaction): 1 μL Round 1 PCR product, 10 μL 2× Taq PCR StarMix with Loading Dye, 1 μL site-specific 3' primer (10 μM), 1 μL L1Hs-nest-28 (10 μM), and PCR-grade water to 20 μL. The cycling program was: 94˚C for 2 min; 35 cycles of 94˚C for 30 sec, 58˚C for 30 sec, 72˚C for 30 sec; 72˚C for 5 min; and held at 4˚C. Validation primer sequences used for each candidate insertion can be found in S10 Table.

All PCR products were run on 2% agarose gels, and gel images were analyzed using Image Lab software (Bio-Rad) to quantify the product sizes. Those PCR products with expected target sizes were cut for downstream analysis. The PCR products were purified with QIAquick Gel Extraction Kit (28706; QIAGEN) and TA-cloned into Trans1-T1 phage-resistant chemically competent cells (Transgen Biotech) using a TOPO TA Cloning Kit for Sequencing (Thermo Fisher Scientific) or the pEASY-T1 Simple Cloning Kit (Transgen Biotech). Positive colony PCR products were Sanger sequenced (Ruibiotech, Inc.) and confirmed (S10 Table).

L1 5' junction nested PCR validation. We provided 22 high-quality step-wise primers covering the full-length L1Hs elements. Each two adjacent step-wise primers were paired and used in 5' junction nested PCR. A total of 100 ng bulk gDNA was used as template for twenty-three 5' junction nested PCR assays (two of them were semi-nested PCR). The PCR protocols were as follows. Round 1 PCR (per reaction): 1 μL diluted DNA, 5 μL KAPA2G Robust HotStart ReadyMix (2×), 0.5 μL site-specific 5' primer (10 μM), 0.5 μL L1Hs step-wise primer (10 μM), and PCR-grade water to 10 μL. The cycling program was: 94˚C for 3 min; 15 cycles of 94˚C for 15 sec, 58˚C for 15 sec, 72˚C for 180 sec; 72˚C for 2 min; and held at 4˚C. Round 2 PCR (per reaction): 1 μL Round 1 PCR product, 10 μL KAPA2G Robust HotStart ReadyMix (2×), 1 μL site-specific 5' primer (10 μM), 1 μL L1Hs step-wise primer (10 μM), and PCR-grade water to 20 μL. The cycling program was: 94˚C for 3 min; 35 cycles of 94˚C for 15 sec, 58˚C for 15 sec, 72˚C for 180 sec; 72˚C for 2 min; and held at 4˚C. Validation primer sequences used for each candidate insertion can be found in S11 Table.

We also successfully obtained the 5' junction of somatic insertion by combining full-length PCR with 5' junction PCR. The PCR protocols were as follows. Round 1 full-length PCR (per reaction): 50 ng DNA, 1 μL site-specific 5' primer (10 μM), 1 μL site-specific 3' primer (10 μM), 0.2 μL Phusion Hot Start II DNA Polymerase (2 U/μL) (Thermo Scientific), 4 μL Phusion HF Buffer (5×), 0.4 μL dNTPs (10 mM), and PCR-grade water to 20 μL. The cycling program was: 98˚C for 3 min; 35 cycles of 98˚C for 10 sec, 60˚C for 15 sec, 72˚C for 240 sec; 72˚C for 4 min; and held at 4˚C. After the full-length PCR, even though none of visible filled-in site band on the agarose gel, we cut the region above the empty site band and performed gel extraction using QIAquick Gel Extraction Kit (28706; QIAGEN). The eluted DNA was used as the template for downstream 5' junction PCR assays. Round 2 PCR (per reaction): 1 μL eluted PCR product, 10 μL KAPA2G Robust HotStart ReadyMix (2×), 1 μL site-specific 5' primer (10 μM), 1 μL L1Hs step-wise primer (10 μM), and PCR-grade water to 20 μL. The cycling program was: 94˚C for 3 min; 35 cycles of 94˚C for 15 sec, 58˚C for 15 sec, 72˚C for 180 sec; 72˚C for 2 min; and held at 4˚C. Validation primer sequences used for each candidate insertion can be found in S10 and S11 Tables.

All PCR products were run on 1.5% agarose gels to quantify the product sizes. Those PCR products with expected target sizes were cut for downstream analysis. The PCR products were purified with QIAquick Gel Extraction Kit (28706; QIAGEN) and TA-cloned into Trans1-T1 phage-resistant chemically competent cells (Transgen Biotech) using a TOPO TA Cloning Kit for Sequencing (Thermo Fisher Scientific) or the pEASY-T1 Simple Cloning Kit (Transgen Biotech). Positive colony PCR products were Sanger sequenced (Genewiz, Inc.) and confirmed (S11 Table).

Sizing of poly-A tail. To investigate poly-A tail length variability, diluted DNA was run through a digital nested PCR (dnPCR) targeting the 3' junction (containing the poly-A tail) of the somatic L1Hs insertion. We used a FAM-labeled primer, L1Hs-nest-28_FAM (10 μM) for Round 2 PCR amplification. dnPCR products were sized by capillary electrophoresis on ABI 3730 DNA Analyzer (Ruibiotech, Inc.) with standard settings for fragment size analysis. Electrophoresis traces were analyzed by GeneMarker software (version: 2.2.0). Sanger sequencing provides the 3' junction amplicon sequence lengths excluding the poly-A tail. For homopolymer, such as poly-A, Sanger sequencing cannot precisely determine the number of bases. Instead, the size of the poly-A tail was calculated as: [measured dnPCR product size]—[known amplicon sequence length excluding the poly-A tail] - 1. The additional 1 bp was subtracted to account for the 3' terminal dA added by Taq [10].

Droplet digital PCR. Custom ddPCR assays were performed with the RainDrop Digital PCR System (RainDance Technologies) following manufacturer's instructions. Each L1Hs assay labeled with FAM was multiplexed with an RNaseP assay labeled with VIC, which served as a genomic copy number reference (copy number = 2) to calculate mosaicism. To confirm the validity of L1Hs assay, we synthesized the L1Hs genome junction oligos of clonal somatic insertion "chr20:2392172". For positive control, we performed multiplexed ddPCR assay using mixed template containing fragmented ACC1 gDNA and diluted synthesized junction. For negative control, we used fragmented ACC1 gDNA, an unrelated adult whose genome do not harbor L1Hs insertion at chr20:2392172. Sequences of primers, probes, and synthesized positive control junction for each of ddPCR assays and droplet statistics can be found in S10 Table.

Supporting information

S1 Fig. Nuclei isolation and NeuN$^+$ fluorescence-activated cell sorting (FACS). (A)–(D) Purify neuronal nuclei from human PFC. (A) The first gate (P1) was set as an FSC-A vs. SSC-A plot to discriminate the population containing small-size debris. (B)–(C) The second (P2) and third (P3) gates were set as FSC-H vs. FSC-W and SSC-H vs. SSC-W plots, respectively, to remove doublets and clumps. (D) Top: NeuN$^-$ and NeuN$^+$ gates were set in the NeuN-PE (561 nm) vs. FSC-A plot. Bottom: a count plot of NeuN-stained nuclei. (E) Purity analysis of sorted neurons. Top: sorted NeuN$^+$ nuclei were re-analyzed by FACS to confirm the sort purity. Bottom: a count plot of re-analyzed NeuN$^+$ nuclei. (F) Purity analysis of sorted glia. Top: sorted NeuN$^-$ nuclei were re-analyzed by FACS to confirm the sort purity. Bottom: a count plot of re-analyzed NeuN$^-$ nuclei. (G) FSC vs. SSC plot of brain homogenate. Brain homogenate contained a huge amount of cell debris and myelin debris. (H) FSC vs. SSC plot of debris-detached single-nuclei homogenate. Minced brain tissue was soaked overnight before homogenization and then incubated with nonionic detergent, Nonidet P-40, to remove cell debris from nuclear membrane. (I) FSC vs. SSC plot of debris removed nuclei fraction. Cell debris and myelin were separated from nuclei using Percoll density gradient

centrifugation. NeuN-PE, PE-conjugated anti-NeuN antibody.
(TIF)

S2 Fig. Confirmation of NeuN$^+$ FACS purity and integrity. (A) Example of fluorescence microscopy confirmation of isolated nuclei. The purity of each fraction was > 95% for NeuN$^+$ and NeuN$^-$ nuclei. Bar = 50 μm. (B)–(C) Examples of integrity confirmation using differential interference contrast (DIC) of sorted neurons (B) and glia (C). Bar = 50 μm. (D) Example of the myelin, lipid, and cell debris layers (12% Percoll) after Percoll density gradient centrifugation. Nuclei were stained with a red fluorescent nuclear counterstain, propidium iodide (PI). Bar = 20 μm. (E) Example of the nuclei fraction layer (35% Percoll) after Percoll density gradient centrifugation. Bar = 50 μm. DAPI, 4',6-diamidino-2-phenylindole; NeuN-PE, PE-conjugated anti-NeuN antibody.
(TIF)

S3 Fig. Schematic diagrams of HAT-seq library construction and computational analysis pipeline. (A) Schematic of the HAT-seq library construction. The fragmented genomic DNA was ligated with P7 truncated adaptors, and then used as template for L1Hs amplification PCR. Primers 1 (P7_Ns_L1Hs) was specific to L1Hs diagnostic "AC" motif. See S1 Table for primer sequences. (B) Schematic of the HAT-seq data analysis pipeline; full details are provided in the Methods.
(TIF)

S4 Fig. EN motif enrichment analysis across all categories of L1Hs insertions. The density distributions of L1 EN motifs around germline KNR (A), UNK (B), somatic insertions (C), randomly sampled positions (D), "Evrony KR" (E), and "Evrony KNR" (F). The lists of "Evrony KR" and "Evrony KNR" were extracted from Evrony *et al*. 2012. The bin size of histogram was 10bp. L1 EN motifs included seven specific motifs (TTAAAA, TTAAGA, TTAGAA, TTG AAA, TTAAAG, CTAAAA, TCAAAA).
(TIF)

S5 Fig. A 5' truncated heart-specific L1Hs insertion (1420_chr10:545758) in a Rett patient. (A) The agarose gel image of 5' junction nested PCR validation for the heart-specific L1Hs insertion in the Rett patient (UMB#1420). The locations of primers used in 5' junction PCR assays were labeled on the top of each lane, where primers with the prime symbol denoted semi-nested PCR assays. The distances between each two adjacent 5' step-wise primers were labeled on the top (dark blue). The yellow line highlighted the expected stair-step bands in 5' junction PCR. 1Kb +: 1 Kb Plus DNA ladder. (B) The Sanger sequencing chromatograms of the 3' and 5' junctions of the somatic insertion (1420_chr10:545758). The L1 EN motif and TSD were indicated by purple and blue lines. (C) Multiple sequence alignment of the 5' end between the identified somatic insertion and three L1Hs consensus sequences (L1Hs Repbase consensus and two hot L1s in human [L1.3 and L1.4]). (D) The schematic structure of the highly 5' truncated (~800 bp) L1Hs insertion 1420_chr10:545758.
(TIF)

S6 Fig. The somatic status of an L1Hs insertions (4516_chr7:14559637) in a Rett patient. The somatic insertion 4516_chr7:14559637 was present in 14 out of 24 nested 3' PCR wells, compared to 24 out of 24 wells for a germline KR insertion (chr7:14629800) from the same donor. DNA sample was diluted to ~300 cells per well. Blue and green arrows indicated bands with target size.
(TIF)

S7 Fig. A full-length embryonic somatic L1Hs insertion (4516_chr20:2392172) in a Rett patient. (A) The agarose gel image of 5' junction nested PCR validation for the embryonic somatic L1Hs insertion (4516_chr20:2392172) in the Rett patient (UMB#4516). The locations of primers used in 5' junction PCR assays were labeled on the top of each lane. Step-wise primers with the prime symbol were used twice in semi-nested PCR assays. The distances between each primer pairs were labeled on the top (dark blue). The yellow line highlighted the expected stair-step bands in 5' junction PCR, while the red lines indicated false positives resulted from non-specific amplification of L1PA subfamilies. 1Kb +: 1 Kb Plus DNA ladder. (B) The Sanger sequencing chromatograms of the 3' and 5' junctions of somatic insertion (4516_chr20:2392 172). The nucleotides shifted chromatogram in 5' junction might result from the DNA polymerase slippage at homopolymers in the upstream region (L1MB3 element), and its sequence was confirmed from the reverse direction. The L1 EN motif and TSD were indicated by purple and blue lines. (C) Multiple sequence alignment of the 5' end between the identified somatic insertion and three L1Hs consensus sequences (L1Hs Repbase consensus and two hot L1s in human [L1.3 and L1.4]). (D) The schematic structure of 4516_chr20:2392172.
(TIF)

S8 Fig. Qualitative and quantitative analysis of L1Hs insertions. (A)–(E) Droplet digital PCR (ddPCR) assays to quantify mosaicism (percentage of cells) of somatic L1Hs insertions at chr20:2392172 in fibroblasts (A) and PFC neurons (B) from Rett patient UMB#4516. Fragmented ACC1 blood gDNA was used as template for negative control assay (C). A mixed template containing fragmented ACC1 blood gDNA and diluted synthesized L1Hs genome junction oligos (D) was used for positive control assay (E). RNaseP served as a genomic copy number reference (copy number = 2). L1Hs and RNaseP assays were labeled with FAM and VIC, respectively. (F)–(G) Relative somatic L1Hs content in PFC neurons and non-brain tissue from the same donor, normalized by the read count of KRs (F) or UNKs (G) from the same tissue sample.
(TIF)

S1 Table. Primer sequences of HAT-seq library.
(XLSX)

S2 Table. ACC1-specific insertions in positive control experiments.
(XLSX)

S3 Table. Statistics of error filters in positive control experiments.
(XLSX)

S4 Table. Clinical characterization of patients with Rett syndrome.
(XLSX)

S5 Table. Statistics of HAT-seq libraries.
(XLSX)

S6 Table. Statistics of known reference insertions among all samples.
(XLSX)

S7 Table. Statistics of polymorphic insertions among all samples.
(XLSX)

S8 Table. Statistics of somatic insertions among all samples.
(XLSX)

S9 Table. TPRT hallmark annotation for all somatic insertions.
(XLSX)

S10 Table. 3' junction nested PCR and digital droplet PCR validation.
(XLSX)

S11 Table. 5' junction nested PCR validation.
(XLSX)

S12 Table. Annotation for somatic exonic insertions.
(XLSX)

S13 Table. Raw data for statistical analyses.
(XLSX)

S14 Table. Quantification statistics among all samples.
(XLSX)

S15 Table. L1Hs enrichment analysis on HAT-seq.
(XLSX)

S16 Table. 3' junction PCR validation for germline insertions.
(XLSX)

S1 Appendix. Cell type-specific sorting for postmortem human brain samples.
(PDF)

S2 Appendix. Identification of ACC1-specific insertions and their zygosity.
(PDF)

S3 Appendix. Detection of somatic insertions in positive control experiments.
(PDF)

S4 Appendix. Benchmarking PCR validation assays for low-frequency somatic insertions.
(PDF)

S5 Appendix. Experimental validation of polymorphic germline L1Hs insertions.
(PDF)

Acknowledgments

We acknowledge the UMB Brain and Tissue Bank (University of Maryland, Baltimore, MD) and the Lieber Institute for Brain Development (Baltimore, MD) for providing postmortem human tissues. We are grateful to Drs Eunjung Alice Lee, Daniel R. Weinberger, Li-Lin Du, Meng-Qiu Dong, Yu Zhang, Ge Gao, Louis Tao, Cheng Li, Jian Lu, and Manyuan Long for their insightful comments and suggestions. We thank Drs Kazuya Iwamoto and Miki Bundo for providing detailed protocols of nuclei isolation. We thank Dr. Timothy W. Yu for providing experimental resources in the revision. We thank the reviewers for constructive feedback on the manuscript. Q-R.L. was supported in part by the Intramural Research Program at the National Institute on Aging, NIH.

Author Contributions

Conceptualization: Boxun Zhao, Qixi Wu, Liping Wei, August Yue Huang.

Data curation: Boxun Zhao, Jing Guo.

Formal analysis: Boxun Zhao, Adam Yongxin Ye.

Funding acquisition: Liping Wei.

Investigation: Boxun Zhao, Qixi Wu, August Yue Huang.

Methodology: Boxun Zhao, Qixi Wu, Jing Guo, Xianing Zheng, Qing-Rong Liu.

Project administration: Boxun Zhao, August Yue Huang.

Resources: Thomas M. Hyde, Liping Wei, August Yue Huang.

Software: Boxun Zhao, Adam Yongxin Ye, Linlin Yan.

Supervision: Liping Wei, August Yue Huang.

Validation: Boxun Zhao, Jing Guo, Xiaoxu Yang.

Visualization: Boxun Zhao, Adam Yongxin Ye, Xiaoxu Yang.

Writing – original draft: Boxun Zhao, Adam Yongxin Ye.

Writing – review & editing: Boxun Zhao, Qixi Wu, Liping Wei, August Yue Huang.

References

1. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. 2015; 31(7):382–92. https://doi.org/10.1016/j.tig.2015.03.013 PMID: 25910407; PubMed Central PMCID: PMCPMC4490042.
2. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. Science. 2013; 341(6141):1237758. https://doi.org/10.1126/science.1237758 PMID: 23828942; PubMed Central PMCID: PMCPMC3909954.
3. Hancks DC, Kazazian HH Jr. Active human retrotransposons: variation and disease. Curr Opin Genet Dev. 2012; 22(3):191–203. https://doi.org/10.1016/j.gde.2012.02.006 PMID: 22406018; PubMed Central PMCID: PMCPMC3376660.
4. Kazazian HH Jr., Moran JV. Mobile DNA in Health and Disease. N Engl J Med. 2017; 377(4):361–70. https://doi.org/10.1056/NEJMra1510092 PMID: 28745987.
5. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993; 72 (4):595–605. PMID: 7679954.
6. Cost GJ, Feng Q, Jacquier A, Boeke JD. Human L1 element target-primed reverse transcription in vitro. EMBO J. 2002; 21(21):5899–910. https://doi.org/10.1093/emboj/cdf592 PMID: 12411507; PubMed Central PMCID: PMCPMC131089.
7. Erwin JA, Paquola AC, Singer T, Gallina I, Novotny M, Quayle C, et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci. 2016; 19(12):1583–91. https://doi.org/10.1038/nn.4388 PMID: 27618310; PubMed Central PMCID: PMCPMC5127747.
8. Muotri AR, Marchetto MC, Coufal NG, Oefner R, Yeo G, Nakashima K, et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010; 468(7322):443–6. https://doi.org/10.1038/nature09544 PMID: 21085180; PubMed Central PMCID: PMCPMC3059197.
9. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009; 460(7259):1127–31. https://doi.org/10.1038/nature08248 PMID: 19657334; PubMed Central PMCID: PMCPMC2909034.
10. Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron. 2015; 85(1):49–59. https://doi.org/10.1016/j.neuron.2014.12.028 PMID: 25569347; PubMed Central PMCID: PMCPMC4299461.
11. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012; 151(3):483–96. https://doi.org/10.1016/j.cell.2012.09.035 PubMed Central PMCID: PMCPMC3567441. PMID: 23101622
12. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sanchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell. 2015; 161(2):228–39. https://doi.org/10.1016/j.cell.2015.03.026 PMID: 25860606; PubMed Central PMCID: PMCPMC4398972.

13. Macia A, Widmann TJ, Heras SR, Ayllon V, Sanchez L, Benkaddour-Boumzaouad M, et al. Engineered LINE-1 retrotransposition in nondividing human neurons. Genome Res. 2017; 27(3):335–48. https://doi.org/10.1101/gr.206805.116 PMID: 27965292; PubMed Central PMCID: PMCPMC5340962.

14. Bundo M, Toyoshima M, Okada Y, Akamatsu W, Ueda J, Nemoto-Miyauchi T, et al. Increased l1 retrotransposition in the neuronal genome in schizophrenia. Neuron. 2014; 81(2):306–13. https://doi.org/10.1016/j.neuron.2013.10.053 PMID: 24389010.

15. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet. 1999; 23(2):185–8. https://doi.org/10.1038/13810 PMID: 10508514.

16. Yu F, Zingler N, Schumann GG, Stratling WH. Mecp2 represses L1 expression and retrotransposition but not Alu transcription. Nucleic Acids Res. 2001; 29(21):4493–501. PMID: 11691937

17. Skene PJ, Illingworth RS, Webb S, Kerr AR, James KD, Turner DJ, et al. Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. Mol Cell. 2010; 37(4):457–68. https://doi.org/10.1016/j.molcel.2010.01.030 PMID: 20188665; PubMed Central PMCID: PMCPMC4338610.

18. Frank SA. Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. Proc Natl Acad Sci U S A. 2010; 107 Suppl 1:1725–30. https://doi.org/10.1073/pnas.0909343106 PMID: 19805033; PubMed Central PMCID: PMCPMC2868288.

19. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015; 25 (10):1499–507. https://doi.org/10.1101/gr.191098.115 PMID: 26430160; PubMed Central PMCID: PMCPMC4579335.

20. Grun D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. Cell. 2015; 163(4):799–810. https://doi.org/10.1016/j.cell.2015.10.039 PMID: 26544934.

21. Faulkner GJ, Garcia-Perez JL. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. Trends Genet. 2017; 33(11):802–16. Epub https://doi.org/10.1016/j.tig.2017.07.004 PMID: 28797643.

22. Evrony GD, Lee E, Park PJ, Walsh CA. Resolving rates of mutation in the brain using single-neuron genomics. Elife. 2016; 5. https://doi.org/10.7554/eLife.12966 PMID: 26901440; PubMed Central PMCID: PMCPMC4805530.

23. Mullen RJ, Buck CR, Smith AM. NeuN, a neuronal specific nuclear protein in vertebrates. Development. 1992; 116(1):201–11. PMID: 1483388.

24. Badge RM, Alisch RS, Moran JV. ATLAS: a system to selectively identify human-specific L1 insertions. Am J Hum Genet. 2003; 72(4):823–38. https://doi.org/10.1086/373939 PMID: 12632328; PubMed Central PMCID: PMCPMC1180347.

25. Ewing AD, Kazazian HH Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010; 20(9):1262–70. https://doi.org/10.1101/gr.106419.110 PMID: 20488934; PubMed Central PMCID: PMCPMC2928504.

26. Tang Z, Steranka JP, Ma S, Grivainis M, Rodic N, Huang CR, et al. Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. Proc Natl Acad Sci U S A. 2017; 114(5):E733–E40. https://doi.org/10.1073/pnas.1619797114 PMID: 28096347; PubMed Central PMCID: PMCPMC5293032.

27. Philippe C, Cristofari G. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. eLife. 2016.

28. Ovchinnikov I, Troxel AB, Swergold GD. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. 2001; 11(12):2050–8. https://doi.org/10.1101/gr.194701 PMID: 11731495; PubMed Central PMCID: PMCPMC311227.

29. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; 6(1):11. https://doi.org/10.1186/s13100-015-0041-9 PMID: 26045719; PubMed Central PMCID: PMCPMC4455052.

30. Mir AA, Philippe C, Cristofari G. euL1db: the European database of L1HS retrotransposon insertions in humans. Nucleic Acids Res. 2015; 43(Database issue):D43–7. https://doi.org/10.1093/nar/gku1043 PMID: 25352549; PubMed Central PMCID: PMCPMC4383891.

31. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2002; 31(2):159–65. https://doi.org/10.1038/ng898 PMID: 12006980.

32. Faulkner GJ, Billon V. L1 retrotransposition in the soma: a field jumping ahead. Mobile DNA. 2018; 9 (1):22. https://doi.org/10.1186/s13100-018-0128-1 PMID: 30002735

33. Grandi FC, Rosser JM, An W. LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. Mol Biol Evol. 2013; 30(3):503–12. https://doi.org/10.1093/molbev/mss251 PubMed Central PMCID: PMCPMC3563966. PMID: 23125228

34. Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature. 2015; 522(7554):89–93. https://doi.org/10.1038/nature14319 PMID: 25762136; PubMed Central PMCID: PMCPMC4480648.

35. King IF, Yandava CN, Mabb AM, Hsiao JS, Huang HS, Pearson BL, et al. Topoisomerases facilitate transcription of long genes linked to autism. Nature. 2013; 501(7465):58–62. https://doi.org/10.1038/nature12504 PMID: 23995680; PubMed Central PMCID: PMCPMC3767287.

36. Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature. 2004; 429(6989):268–74. https://doi.org/10.1038/nature02536 PMID: 15152245.

37. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 1999; 9(6):657–63. https://doi.org/10.1016/S0959-437X(99)00031-3 PMID: 10607616

38. Guo C, Jeong H-H, Hsieh Y-C, Klein H-U, Bennett DA, Jager PL, et al. Tau Activates Transposable Elements in Alzheimer's Disease. Cell reports. 2018; 23(10):2874–80. https://doi.org/10.1016/j.celrep.2018.05.004 PMID: 29874575

39. Thomas H, Beck K, Adamczyk M, Aeschlimann P, Langley M, Oita RC, et al. Transglutaminase 6: a protein associated with central nervous system development and motor function. Amino Acids. 2013; 44 (1):161–77. https://doi.org/10.1007/s00726-011-1091-z PMID: 21984379; PubMed Central PMCID: PMCPMC3535377.

40. Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain. 2010; 133(Pt 12):3510–8. https://doi.org/10.1093/brain/awq323 PMID: 21106500.

41. Li M, Pang SY, Song Y, Kung MH, Ho SL, Sham PC. Whole exome sequencing identifies a novel mutation in the transglutaminase 6 gene for spinocerebellar ataxia in a Chinese family. Clin Genet. 2013; 83 (3):269–73. https://doi.org/10.1111/j.1399-0004.2012.01895.x PMID: 22554020.

42. Guo YC, Lin JJ, Liao YC, Tsai PC, Lee YC, Soong BW. Spinocerebellar ataxia 35: novel mutations in TGM6 with clinical and genetic characterization. Neurology. 2014; 83(17):1554–61. https://doi.org/10.1212/WNL.0000000000000909 PMID: 25253745.

43. Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol. 2001; 18(6):926–35. https://doi.org/10.1093/oxfordjournals.molbev.a003893 PMID: 11371580

44. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, et al. Human l1 retrotransposition is associated with genetic instability in vivo. Cell. 2002; 110(3):327–38. PMID: 12176320

45. Gilbert N, Lutz S, Morrish TA, Moran JV. Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol. 2005; 25(17):7780–95. https://doi.org/10.1128/MCB.25.17.7780-7795.2005 PMID: 16107723

46. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, et al. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 2012; 22(12):2328–38. https://doi.org/10.1101/gr.145235.112 PMID: 22968929; PubMed Central PMCID: PMCPMC3514663.

47. Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, et al. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res. 2015; 25(10):1536–45. https://doi.org/10.1101/gr.196238.115 PMID: 26260970; PubMed Central PMCID: PMCPMC4579339.

48. Doucet-O'Hare TT, Rodic N, Sharma R, Darbari I, Abril G, Choi JA, et al. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc Natl Acad Sci U S A. 2015; 112 (35):E4894–900. https://doi.org/10.1073/pnas.1502474112 PMID: 26283398; PubMed Central PMCID: PMCPMC4568228.

49. Achanta P, Steranka JP, Tang Z, Rodic N, Sharma R, Yang WR, et al. Somatic retrotransposition is infrequent in glioblastomas. Mob DNA. 2016; 7:22. https://doi.org/10.1186/s13100-016-0077-5 PMID: 27843500; PubMed Central PMCID: PMCPMC5105304.

50. Reilly MT, Faulkner GJ, Dubnau J, Ponomarev I, Gage FH. The role of transposable elements in health and diseases of the central nervous system. J Neurosci. 2013; 33(45):17577–86. https://doi.org/10.1523/JNEUROSCI.3369-13.2013 PMID: 24198348; PubMed Central PMCID: PMCPMC3818539.

51. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci. 2014; 15(8):497–506. https://doi.org/10.1038/nrn3730 PMID: 25005482; PubMed Central PMCID: PMCPMC4443810.

52. Goodier JL. Retrotransposition in tumors and brains. Mob DNA. 2014; 5(1):11. https://doi.org/10.1186/1759-8753-5-11 PMID: 24708615; PubMed Central PMCID: PMCPMC3995500.

53. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A. 1997; 94(5):1872–7. https://doi.org/10.1073/pnas.94.5.1872 PMID: 9050872; PubMed Central PMCID: PMCPMC20010.