1 **Title**

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

4 **Authors and Affiliations**

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

7 1 National Institute of Biological Sciences, Beijing, 102206, China

8 2 Graduate School of Peking Union Medical College, Beijing, 100730, China

9 3 School of Life Sciences, Peking University, Beijing, 100871, China

10 4 Peking-Tsinghua Center for Life Sciences, Beijing, 100871, China

11 5 Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life
12 Sciences, Peking University, Beijing, 100871, China

13 6 Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

14 7 College of Life Sciences, Beijing Normal University, Beijing, 100875, China

15 8 Laboratory of Clinical Investigation, National Institute on Aging, Baltimore, MD 21224, USA

16 9 Lieber Institute for Brain Development, Baltimore, MD 21205, USA

17 10 Departments of Psychiatry & Behavioral Sciences and Neurology, Johns Hopkins University
18 School of Medicine, Baltimore, MD 21205, USA

19 *Correspondence

Correspondence should be addressed to A.Y.H. (huangy@mail.cbi.pku.edu.cn) and L.W. (weilp@mail.cbi.pku.edu.cn).

Short title

Somatic L1 mosaicism in Rett and healthy individuals

Keywords

mosaicism, retrotransposition, somatic insertion, Rett syndrome

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 L1Hs (Human-specific LINE-1) 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 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 (Campbell et al., 2015). In addition to various types of cancers, somatic mosaicisms reportedly contribute to a variety of neurological disorders, including epilepsy, neurodegeneration, and hemimegalencephaly (Poduri et al., 2013). The human-specific LINE-1 (L1Hs) retrotransposon family is the only known family of active autonomous transposons in the human genome (Hancks and Kazazian, 2012; Kazazian and Moran, 2017). 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 (Cost et al., 2002; Luan et al., 1993). Mounting evidence supports that L1Hs elements, with increased copy number in the brain relative to

other tissues, contribute to neuronal diversity via somatic retrotransposition (Coufal et al., 2009; Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015; Muotri et al., 2010; Upton et al., 2015).

Recent studies reported the occurrence of somatic L1Hs insertions during neurogenesis and in non-dividing mature neurons (Coufal et al., 2009; Macia et al., 2017). Other studies have observed dysregulated L1Hs copy number in patients with Rett syndrome (Muotri et al., 2010) and schizophrenia (Bundo et al., 2014). Methyl-CpG binding protein 2 (*MECP2*) is the major disease-causing gene of Rett syndrome (Amir et al., 1999). Its gene product, MeCP2, can bind to the 5' UTR of L1 elements and represses their expression and retrotransposition (Yu et al., 2001). 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 (Muotri et al., 2010; Skene et al., 2010), 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 (Frank, 2010). Single-cell targeted sequencing approaches have been used to identify somatic insertions (Erwin et al., 2016; Evrony et al., 2012; Upton et al., 2015). However, such methods typically require a large number of cells and demand considerable sequencing depth for unbiased profiling of human tissues (Grun and van Oudenaarden, 2015; Navin, 2015). 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 (Navin, 2015).

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 (Faulkner and Garcia-Perez, 2017). 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 (Evrony et al., 2016).

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

103 insertion signals and specific identification of true signals from background noise. Enriching neuronal
104 nuclei from bulk brain tissue facilitates the accurate deciphering of cell type-specific characteristic and
105 increases the chance of identifying clonal somatic insertions that are derived from the same progenitor
106 cell and shared by multiple neurons. Therefore, we labeled prefrontal cortex (PFC) neuronal nuclei
107 using an antibody against neuron-specific marker NeuN (Mullen et al., 1992), and subsequently
108 purified $NeuN^+$ nuclei from postmortem human PFC by fluorescence-activated cell sorting (FACS)
109 (Fig 1A; S1A–D Fig; Appendix 1). All initially sorted nuclei were re-analyzed with a second round of
110 FACS, and the purity of the initial sorting was found to be > 96% (S1E–F Fig; Appendix 1). The
111 integrity and purity of sorted nuclei were confirmed by fluorescence microscopy (S2A–C Fig).
112
113 To distinguish the signals of active L1Hs elements from other transposon families that are typically
114 inactive in human, we developed a method called human active transposon sequencing (HAT-seq) (Fig
115 1B; S3A Fig; S1 Table) based on ATLAS (Badge et al., 2003) and several versions of high-throughput
116 sequencing-based L1 amplification methods (Erwin et al., 2016; Ewing and Kazazian, 2010; Philippe
117 and Cristofari, 2016; Tang et al., 2017). Firstly, L1Hs insertions were specifically enriched and
118 amplified using a primer targeting the diagnostic “AC” motif of L1Hs (Hancks and Kazazian, 2012;
119 Ovchinnikov et al., 2001). To ameliorate the poor performance of Illumina sequencing platform for
120 low-diversity libraries, we employed a nucleotides-shifting design by adding two, four, or six random
121 nucleotides upstream of the L1Hs-specific primer, which greatly increased the diversity of the
122 structure-transformed semi-amplicon library and markedly improved the sequencing quality of L1Hs
123 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.

144 **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) (Bao et al., 2015).
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 (Upton et al., 2015). Read was re-aligned to hg19 using BLAST to find the corresponding best alignments for the non-retrotransposon and retrotransposon segments. Read was removed as a putative chimera when the overlap of the two best segments was > 10 bp and A% ≥ 50% or 6–10 bp and 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 (Mir et al., 2015).
Misaligned reads	We rejected reads at risk of being misaligned, defined as inconsistent BWA and BLAT alignment.
Local SV	We rejected reads at risk of being derived from a nearby reference L1Hs (Upton et al., 2015). We extracted 2 kb from the reference genome extending downstream from an aligned non-retrotransposon section and aligned the full read contig against this region with BLAT to exclude genomic rearrangements.
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.

145

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

147 To benchmark the performance of HAT-seq for detecting somatic L1Hs insertions, we experimentally
148 generated a series of positive control samples with insertions at different frequencies by mixing the
149 genomic DNA (gDNA) extracted from the blood samples of two unrelated adults, ACC1 and ACC2
150 (see details in Materials and Methods). 172 ACC1 non-reference germline L1Hs insertions were
151 identified by HAT-seq, 64 of which were confirmed to be ACC1-specific by 3' junction PCR (3' PCR)
152 analysis of gDNA from ACC1 and ACC2 (Fig 2A; S2 Table; Appendix 2) and thus served as positive
153 controls. Three HAT-seq libraries were generated from samples consisting of ACC2 gDNA spiked with
154 1%, 0.1%, or 0.01% of ACC1 gDNA. Considering that decreasing the number of cells pooled for
155 sequencing increased the signal-to-noise ratio for detecting somatic insertions (Evrony et al., 2016),
156 each HAT-seq library was constructed from 20 ng input (about 3,000 cells).
157
158 The zygosity of ACC1-specific L1Hs insertions was confirmed by full-length PCR: 49 of which were
159 heterozygous, 9 of which were homozygous, and 6 of which were zygosity-undetermined (Fig 2B; S2
160 Table; Appendix 2). We detected all 64 ACC1-specific insertions in our positive control 1% ACC1
161 spike-in library, 49 (76.6%) of which passed all of error filters and subsequently were deemed
162 "identified" by HAT-seq. In the 0.1% library, we detected 23 ACC1-specific insertions (16
163 heterozygous, 4 homozygous, and 3 zygosity-undetermined), 17 (73.9%) of which were identified. In
164 the 0.01% library, we detected seven heterozygous ACC1-specific insertions, five (71.4%) of which
165 were identified. The distributions of signal counts (reads with unique start positions) per ACC1-
166 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 Materials and 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 Appendix 3).

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 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 Table). 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%. 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 Materials and 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 (Morrish et al., 2002). These features of somatic L1Hs insertions helped to elucidate the specificity of HAT-seq method.

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	Fibro blast	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).

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 (Appendix 4). In theory, if a somatic insertion was unique to a single cell, it would be impossible to detect it 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 (Evrony et al., 2015) 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–H 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 (Faulkner and Billon, 2018). 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; Appendix 4). 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 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–D). Notably, we also validated this 5' junction by combining full-length PCR with 5' junction PCR (Fig 4E; see details in Materials and Methods).

In addition, we verified one fibroblast- and another heart-specific L1Hs insertion in two patients with Rett syndrome (Fig 3E–F). 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 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 (Evrony et al., 2015; Grandi et al., 2013), 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 both PFC neurons, PFC glia, and fibroblasts (Fig 3G–H 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–E Fig and S10 Table). Our observations demonstrated that endogenous L1Hs could retrotranspose in various types of non-brain tissues during human development.

252 **Abnormal L1Hs mobilization in patients with Rett syndrome**

253 Our HAT-seq bulk sequencing data enabled us to perform statistical analysis of the exonic and intronic
254 patterns of somatic L1Hs insertions in samples from Rett patients and matched healthy controls. We
255 found 180 somatic insertions that were integrated into exonic regions: 9 of which were located in 5'
256 UTR, 102 of which were located in coding regions, and 69 of which were located in 3' UTR (S12
257 Table). While no significant difference was observed in introns (OR = 0.97, p = 0.44, Fisher's exact
258 test), somatic insertions were significantly depleted in exons (OR = 0.59, $p = 6.6\times10^{-4}$, Fisher's exact
259 test) of Rett patients compared with matched healthy controls (Fig 5A and S13 Table). Previous studies
260 have shown that dysregulation of long genes (> 100 kb) was linked to neurological disorders, including
261 Rett syndrome (Gabel et al., 2015) and autism spectrum disorder (King et al., 2013). We used our HAT-
262 seq data to investigate somatic insertional bias in both long (> 100 kb) and short genes (< 100 kb) of
263 Rett patients. As a result, we found significant depletion of somatic insertions in exons of long genes
264 (OR = 0.27, $p = 5.2\times10^{-5}$, Fisher's exact test) but not short genes (OR = 0.76, p = 0.12, Fisher's exact
265 test; Fig 5B and S13 Table). Our speculation was that if an L1Hs inserted into the exonic regions,
266 especially in important genes, of the *MECP2* mutated cell, the cell would have a higher risk of death
267 and subsequently be cleared up; therefore, the observed exonic depletion of L1 insertions in Rett
268 patients might be resulted from the negative selection acting on those "lethal" exonic insertions.
269
270 In contrast to germline insertions, the impact of somatic insertions depends not only on their genomic
271 location, but also the number of cells carrying that insertion, highlighting the importance of clonal
272 somatic insertions. We found that in cortical neurons of Rett patients, clonal somatic insertions were
273 enriched in introns (OR = 1.85, p = 0.029, Fisher's exact test; Fig 5C and S13 Table); these clonal

274 intronic insertions were significantly enriched in the sense orientation to the transcripts (OR = 3.3, p
275 = 0.0067, Fisher's exact test; Fig 5D and S13 Table). The presence of L1 insertion in the sense
276 orientation has been reported to interfere with transcriptional elongation of co-localized genes (Han et
277 al., 2004). Considering that clonal insertions are more likely to have occurred at an early stage of
278 development and thus affect a relatively large proportion of cells, these distinct insertion pattern in
279 cortical neurons of Rett patients might indicate potential transcriptional burden on the nervous system.
280

281 **Genomic patterns of somatic and germline L1Hs insertions**

282 The design of HAT-seq method allowed for unbiased enrichment of both somatic and germline L1Hs
283 insertions from each of bulk DNA samples. As germline insertion had constant genomic copy number
284 in all tissues from the same donor, we used germline insertion as endogenous control to measure the
285 relative copy number of genome-wide somatic insertions in the brain and non-brain tissues. We
286 quantified the relative somatic L1Hs content by calculating the L1Hs-derived read count ratio of
287 somatic to germline insertions using HAT-seq data of each sample (S14 Table; see details in Materials
288 and Methods). Among all Rett patients and their matched controls, we observed a significant increase
289 in the copy number of somatic L1Hs insertions in PFC neurons relative to matched non-brain tissues
290 (heart, eye, or fibroblast) from the same donor (n = 10, $p = 2.7\times10^{-4}$, paired *t*-test; Fig 6A and S8F–G
291 Fig). We also estimated the occurrence rate of somatic L1Hs insertions based on the germline insertion
292 copy number of each individual (Fig 6B). This produced an average of 1.29 [95% CI: 1.03–1.55]
293 somatic insertions per PFC neuron versus 0.60 [95% CI: 0.46–0.74] insertions per non-brain cell (S14
294 Table). Our observation of higher somatic L1Hs rate in PFC neurons from healthy individuals argued

295 for the active retrotransposition of L1Hs in the human brain (Coufal et al., 2009). One significant
296 advantage of HAT-seq was the ability to distinguish signals of somatic insertions from the
297 overwhelming copies of germline L1Hs insertions in the genome (see details in Materials and
298 Methods). Inconsistent with the previous qPCR result (Muotri et al., 2010), when comparing the group
299 of Rett patients with matched healthy controls, we only observed a slight but not significant increase
300 of somatic L1Hs insertion rate in the Rett group, with 1.36 [range: 0.89–1.82] versus 1.22 [range:
301 0.63–1.66] per PFC neuron and 0.66 [range: 0.37–1.01] versus 0.54 [range: 0.34–0.72] per non-brain
302 cell (Fig 6C and S14 Table).
303
304 We next characterized the genome-wide germline L1Hs insertions. HAT-seq yielded greater than 320-
305 fold enrichment for KR, KNR, and UNK L1Hs insertions (S15 Table). On average, 814 KRs, 183
306 KNRs, and 10 UNKs were identified in each bulk sample (Table 2, S5–7 Table). Hierarchical clustering
307 based on L1Hs profiles correctly paired all neuronal samples with the non-neuronal tissue samples of
308 the same individual (Fig 6D). To experimentally validate the HAT-seq predicted germline insertions,
309 we performed 3' PCR validation on a random subset of polymorphic insertions from among the ten
310 individuals, including 8 sites out of 160 polymorphic KRs, 20 sites out of 451 KNRs, and 2 sites out
311 of 48 UNKs (S7 and S16 Table). As a result, all of the assayed sites were detected in 3' PCR, with
312 98.4% (120/122) and 100% (168/168) sensitivity and specificity, respectively (S16 Table and
313 Appendix 5). These results support that HAT-seq can reliably detect germline L1Hs insertions with
314 high sensitivity and specificity.
315

316 Previous studies have shown that intronic germline L1Hs insertions are sense-depleted (Ewing and
317 Kazazian, 2010; Smit, 1999; Upton et al., 2015). As expected, the germline insertions identified in this
318 study were significantly sense-depleted to the transcripts (633/1,544 [41%], $p = 1.6\times10^{-12}$, binomial
319 test; Fig 6E and S13 Table). It is important to ask the question: whether such orientation bias for
320 germline insertions is resulted from natural selection or insertional preference? To address this, we
321 chose somatic L1Hs insertions as internal reference to control confounding factors. We compared the
322 orientation bias between germline and somatic L1Hs insertions in transcripts and found that germline
323 insertions were significantly sense-depleted than somatic insertions (odds ratio [OR] = 0.79, $p =$
324 7.9×10^{-4}, Fisher's exact test; Fig 6E and S13 Table). Because somatic L1Hs insertions only affected a
325 small proportion of cells and thus they should undergo weaker selective pressure than germline
326 insertions, our results suggested that natural selection may play a major role in shaping the sense-
327 depleted distribution of germline L1Hs insertions.
328

329 **Discussion**

330 Here, we present HAT-seq, a bulk DNA sequencing method to profile genome-wide L1Hs insertions
331 from physiologically normal and pathological human tissues. We demonstrated that, in addition to
332 neuronal cells (Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015; Macia et al., 2017; Upton
333 et al., 2015), L1Hs also retrotransposed in a variety of non-brain tissues and cell types during normal
334 development and contributed to the inter-cellular diversity of the human genome. Using high-
335 throughput sequencing-based quantitative analysis, we found that somatic insertions occurred at a
336 higher rate in brain than in non-brain tissues, consistent with previous studies (Coufal et al., 2009).

337 Previous qPCR and single-cell genomic studies have resulted in conflicting estimates of the frequency
338 of somatic insertions in neurons: ~80 L1 insertions per neuron (Coufal et al., 2009), < 0.04–0.6 L1
339 insertions per neuron (Evrony et al., 2012), 13.7 L1 insertions per neuron (Upton et al., 2015), or
340 ~0.58–1 somatic L1-associated variants per neuron (Erwin et al., 2016). Differential estimates might
341 result from differences in WGA and signal enrichment methods. Using a bulk DNA sequencing
342 approach, we estimated the rate of somatic insertions to be 0.63–1.66 L1Hs insertions per PFC neuron
343 in healthy individuals (Fig 6B and S14 Table).
344
345 Clonally distributed insertions are prevalent in normal brain (Evrony et al., 2015). Increasing evidence
346 suggests that neuronal L1s retrotransposition contributes to the susceptibility to and pathophysiology
347 of neurological disorders, including Rett syndrome (Muotri et al., 2010), schizophrenia (Bundo et al.,
348 2014) and Alzheimer's disease (Guo et al., 2018). We observed that, in PFC neurons of Rett patients,
349 clonal somatic insertions were enriched in introns, and these clonal intronic insertions were
350 significantly enriched in the sense orientation (Fig 5C–D). In particular, in Rett patient UMB#4516,
351 we found a full-length, sense-orientated, intronic somatic insertion (chr20:2392172) in *TGM6* (S7 Fig
352 and S11 Table), a gene associated with central nervous system development and motor function
353 (Thomas et al., 2013), which could potentially dysregulate gene expression (Han et al., 2004). We
354 found that 6.34% of fibroblasts and 2.87% of PFC neurons contained this insertion (S8A–E Fig and
355 S10 Table), suggesting that it might occur in the 16-cell or 32-cell stages during morula stage.
356 Mutations in *TGM6* are associated with spinocerebellar ataxia type 35, one of a group of genetic
357 disorders characterized by poor coordination of hands, gait, speech, and eye movements as well as
358 frequent atrophy of the cerebellum (Guo et al., 2014; Li et al., 2013; Wang et al., 2010). According to

359 the clinical records, UMB#4516 had slight cerebral atrophy and cerebellar degeneration, could not
360 hold things in her hands, and her speech development ceased at 16 months of age; these phenotypes
361 were absent in the other four patients with Rett syndrome. Taken together, our data indicated that this
362 clonal L1Hs insertion of *TGM6* might be correlated with the distinct clinical phenotype of UMB#4516.
363
364 Previous studies have provided evidence for significant selection against older L1 elements that are
365 non-polymorphic (Boissinot et al., 2001; Ewing and Kazazian, 2010). To characterize the insertion
366 pattern of L1 with minimal influence from selective pressure, experimental methods were developed
367 for recovery of novel L1 insertions in cultured cells (Gilbert et al., 2005; Symer et al., 2002). Using
368 HAT-seq method, we were able to distinguish somatic L1Hs insertions from germline L1Hs insertions
369 within the same individual. To determine whether the sense-depleted germline insertion was resulted
370 from natural selection or insertional preference, we used somatic insertion as internal reference to
371 control confounding factors such as intrinsic insertion preference and compared germline with somatic
372 insertions. Our results suggested that natural selection shaped a sense-depleted distribution of germline
373 L1Hs insertions in the human genome.
374
375 Several PCR-based bulk sequencing methods, such as ATLAS (Badge et al., 2003), L1-seq (Ewing
376 and Kazazian, 2010), TIP-seq (Tang et al., 2017), bulk SLAV-seq (Erwin et al., 2016), and ATLAS-seq
377 (Philippe and Cristofari, 2016), have been developed to identify germline L1Hs insertions.
378 Furthermore, L1-seq and TIP-seq have been successfully used in the identification of somatic
379 insertions in tumors (Achanta et al., 2016; Doucet-O'Hare et al., 2015; Ewing et al., 2015; Solyom et
380 al., 2012; Tang et al., 2017). 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 (Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015; Upton et al., 2015), 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 (Evrony et al., 2016). 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 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

403 the key filter "chimera within poly-A tail" was not applicable for unassembled read pairs, our sequence
404 analysis suggested that chimeric molecules bridging within the poly-A tail was the major source of
405 false-positives for unassembled data (see details in Materials and Methods). As shown in the statistics
406 of positive control libraries (S2 Table) and experimental validation, the unassembled data could
407 provide additional signals of somatic L1Hs insertions but require careful analysis and rigorous
408 validation to address technical artifacts. Further gains in statistical power will be benefited from
409 increased sample size and improved efficiency of HAT-seq.

410

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

419

420 **Acknowledgements**

421 We acknowledge the UMB Brain and Tissue Bank (University of Maryland, Baltimore, MD) and the
422 Lieber Institute for Brain Development (Baltimore, MD) for providing postmortem human tissues. We
423 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. This study was supported by the National Natural Science Foundation of China (31530092) and the Ministry of Science and Technology 863 Grant (2015AA020108). Q-R.L. was supported in part by the Intramural Research Program at the National Institute on Aging.

Author contributions

Adam Yongxin Ye

Formal analysis, Methodology, Visualization, Writing-original draft

August Yue Huang

Conceptualization, Resources, Supervision, Writing—review and editing

Boxun Zhao

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing—original draft, Writing—review and editing

Jing Guo

Data curation, Methodology, Validation

446 Liping Wei

447 Conceptualization, Funding acquisition, Resources, Supervision, Writing—review and editing

448

449 Linlin Yan

450 Formal analysis

451

452 Thomas M. Hyde

453 Resources

454

455 Qing-Rong Liu

456 Methodology

457

458 Qixi Wu

459 Conceptualization, Investigation, Methodology, Writing—review and editing

460

461 Xianing Zheng

462 Methodology

463

464 Xiaoxu Yang

465 Validation, Visualization

466

Competing interests

No conflicts of interest.

References

Achanta, P., Steranka, J.P., Tang, Z., Rodic, N., Sharma, R., Yang, W.R., Ma, S., Grivainis, M., Huang, C.R.L., Schneider, A.M., *et al.* (2016). Somatic retrotransposition is infrequent in glioblastomas. Mob DNA *7*, 22.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet *23*, 185-188.

Badge, R.M., Alisch, R.S., and Moran, J.V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. Am J Hum Genet *72*, 823-838.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA *6*, 11.

Boissinot, S., Entezam, A., and Furano, A.V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol *18*, 926-935.

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., Sunaga, F., Toritsuka, M., Ikawa, D., Kakita, A., *et al.* (2014). Increased l1 retrotransposition in the neuronal genome in schizophrenia. Neuron *81*, 306-313.

Campbell, I.M., Shaw, C.A., Stankiewicz, P., and Lupski, J.R. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends Genet *31*, 382-392.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. EMBO J *21*, 5899-5910.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127-1131.

Doucet-O'Hare, T.T., Rodic, N., Sharma, R., Darbari, I., Abril, G., Choi, J.A., Young Ahn, J., Cheng, Y., Anders, R.A., Burns, K.H., *et al.* (2015). LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc Natl Acad Sci U S A *112*, E4894-4900.

Erwin, J.A., Marchetto, M.C., and Gage, F.H. (2014). Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci *15*, 497-506.

Erwin, J.A., Paquola, A.C., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I., Butcher, C.R., Herdy, J.R., *et al.* (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci *19*, 1583-1591.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., *et al.* (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J., *et al.* (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49-59.

Evrony, G.D., Lee, E., Park, P.J., and Walsh, C.A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. Elife *5*.

Ewing, A.D., Gacita, A., Wood, L.D., Ma, F., Xing, D., Kim, M.S., Manda, S.S., Abril, G., Pereira, G., Makohon-Moore, A., *et al.* (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res *25*, 1536-1545.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

Faulkner, G.J., and Billon, V. (2018). L1 retrotransposition in the soma: a field jumping ahead. Mobile DNA *9*, 22.

Faulkner, G.J., and Garcia-Perez, J.L. (2017). L1 Mosaicism in Mammals: Extent, Effects, and Evolution. Trends Genet *33*, 802-816.

Frank, S.A. (2010). 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 *107 Suppl 1*, 1725-1730.

Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature *522*, 89-93.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition

intermediates in cultured human cells. Mol Cell Biol *25*, 7780-7795.

Goodier, J.L. (2014). Retrotransposition in tumors and brains. Mob DNA *5*, 11.

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

Grun, D., and van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing Experiments. Cell *163*, 799-810.

Guo, C., Jeong, H.-H., Hsieh, Y.-C., Klein, H.-U., Bennett, D.A., Jager, P.L., Liu, Z., and Shulman, J.M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. Cell reports *23*, 2874-2880.

Guo, Y.C., Lin, J.J., Liao, Y.C., Tsai, P.C., Lee, Y.C., and Soong, B.W. (2014). Spinocerebellar ataxia 35: novel mutations in TGM6 with clinical and genetic characterization. Neurology *83*, 1554-1561.

Han, J.S., Szak, S.T., and Boeke, J.D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature *429*, 268-274.

Hancks, D.C., and Kazazian, H.H., Jr. (2012). Active human retrotransposons: variation and disease. Curr Opin Genet Dev *22*, 191-203.

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

Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N Engl J Med *377*, 361-370.

King, I.F., Yandava, C.N., Mabb, A.M., Hsiao, J.S., Huang, H.S., Pearson, B.L., Calabrese, J.M., Starmer, J., Parker, J.S., Magnuson, T.*, et al.* (2013). Topoisomerases facilitate transcription of long genes linked to autism. Nature *501*, 58-62.

Li, M., Pang, S.Y., Song, Y., Kung, M.H., Ho, S.L., and Sham, P.C. (2013). Whole exome sequencing identifies a novel mutation in the transglutaminase 6 gene for spinocerebellar ataxia in a Chinese family. Clin Genet *83*, 269-273.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell *72*, 595-605.

Macia, A., Widmann, T.J., Heras, S.R., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., Munoz-Lopez, M., Rubio, A., Amador-Cubero, S., Blanco-Jimenez, E.*, et al.* (2017). Engineered LINE-1

553 retrotransposition in nondividing human neurons. Genome Res *27*, 335-348.

554 Mir, A.A., Philippe, C., and Cristofari, G. (2015). euL1db: the European database of L1HS
555 retrotransposon insertions in humans. Nucleic Acids Res *43*, D43-47.

556 Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and
557 Moran, J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition.
558 Nat Genet *31*, 159-165.

559 Mullen, R.J., Buck, C.R., and Smith, A.M. (1992). NeuN, a neuronal specific nuclear protein in
560 vertebrates. Development *116*, 201-211.

561 Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H.
562 (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature *468*, 443-446.

563 Navin, N.E. (2015). The first five years of single-cell cancer genomics and beyond. Genome Res *25*,
564 1499-1507.

565 Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human
566 LINE-1 insertions: evidence supporting random insertion. Genome Res *11*, 2050-2058.

567 Philippe, C., and Cristofari, G. (2016). Activation of individual L1 retrotransposon instances is
568 restricted to cell-type dependent permissive loci. eLife.

569 Poduri, A., Evrony, G.D., Cai, X., and Walsh, C.A. (2013). Somatic mutation, genomic variation, and
570 neurological disease. Science *341*, 1237758.

571 Reilly, M.T., Faulkner, G.J., Dubnau, J., Ponomarev, I., and Gage, F.H. (2013). The role of transposable
572 elements in health and diseases of the central nervous system. J Neurosci *33*, 17577-17586.

573 Skene, P.J., Illingworth, R.S., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Andrews, R., and Bird,
574 A.P. (2010). Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the
575 chromatin state. Mol Cell *37*, 457-468.

576 Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian
577 genomes. Curr Opin Genet Dev *9*, 657-663.

578 Solyom, S., Ewing, A.D., Rahrmann, E.P., Doucet, T., Nelson, H.H., Burns, M.B., Harris, R.S., Sigmon,
579 D.F., Casella, A., Erlanger, B.*, et al.* (2012). Extensive somatic L1 retrotransposition in colorectal
580 tumors. Genome Res *22*, 2328-2338.

581 Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002).

582 Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

583 Tang, Z., Steranka, J.P., Ma, S., Grivainis, M., Rodic, N., Huang, C.R., Shih, I.M., Wang, T.L., Boeke,
584 J.D., Fenyo, D.*, et al.* (2017). Human transposon insertion profiling: Analysis, visualization and
585 identification of somatic LINE-1 insertions in ovarian cancer. Proc Natl Acad Sci U S A *114*, E733-
586 E740.

587 Thomas, H., Beck, K., Adamczyk, M., Aeschlimann, P., Langley, M., Oita, R.C., Thiebach, L., Hils,
588 M., and Aeschlimann, D. (2013). Transglutaminase 6: a protein associated with central nervous system
589 development and motor function. Amino Acids *44*, 161-177.

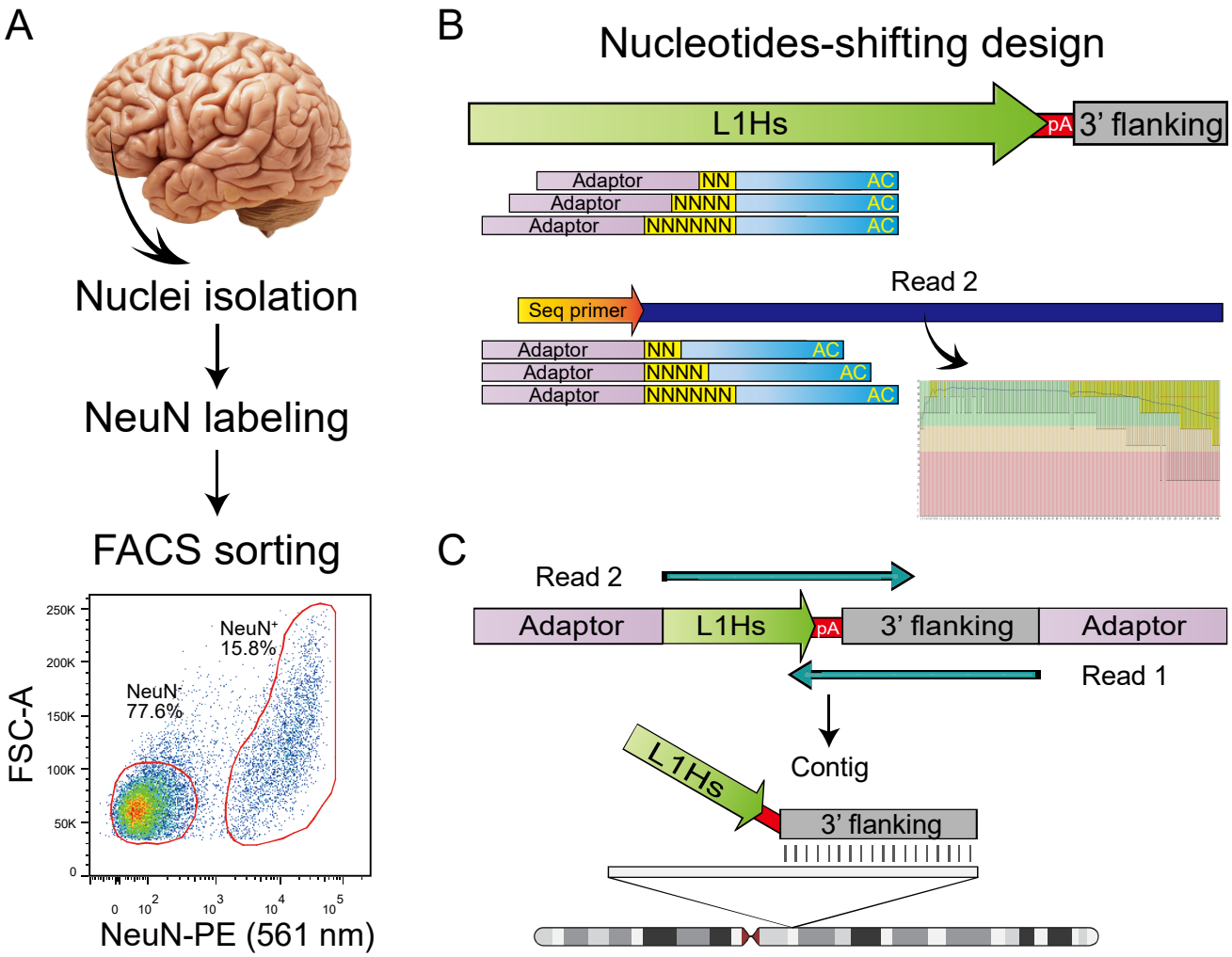
590 Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O.,
591 Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M.*, et al.* (2015). Ubiquitous
592 L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

593 Wang, J.L., Yang, X., Xia, K., Hu, Z.M., Weng, L., Jin, X., Jiang, H., Zhang, P., Shen, L., Guo, J.F.*, et*
594 *al.* (2010). TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome
595 sequencing. Brain *133*, 3510-3518.

596 Yu, F., Zingler, N., Schumann, G.G., and Stratling, W.H. (2001). Mecp2 represses L1 expression and
597 retrotransposition but not Alu transcription. Nucleic Acids Res *29*, 4493-4501.

598

599 **Figures**

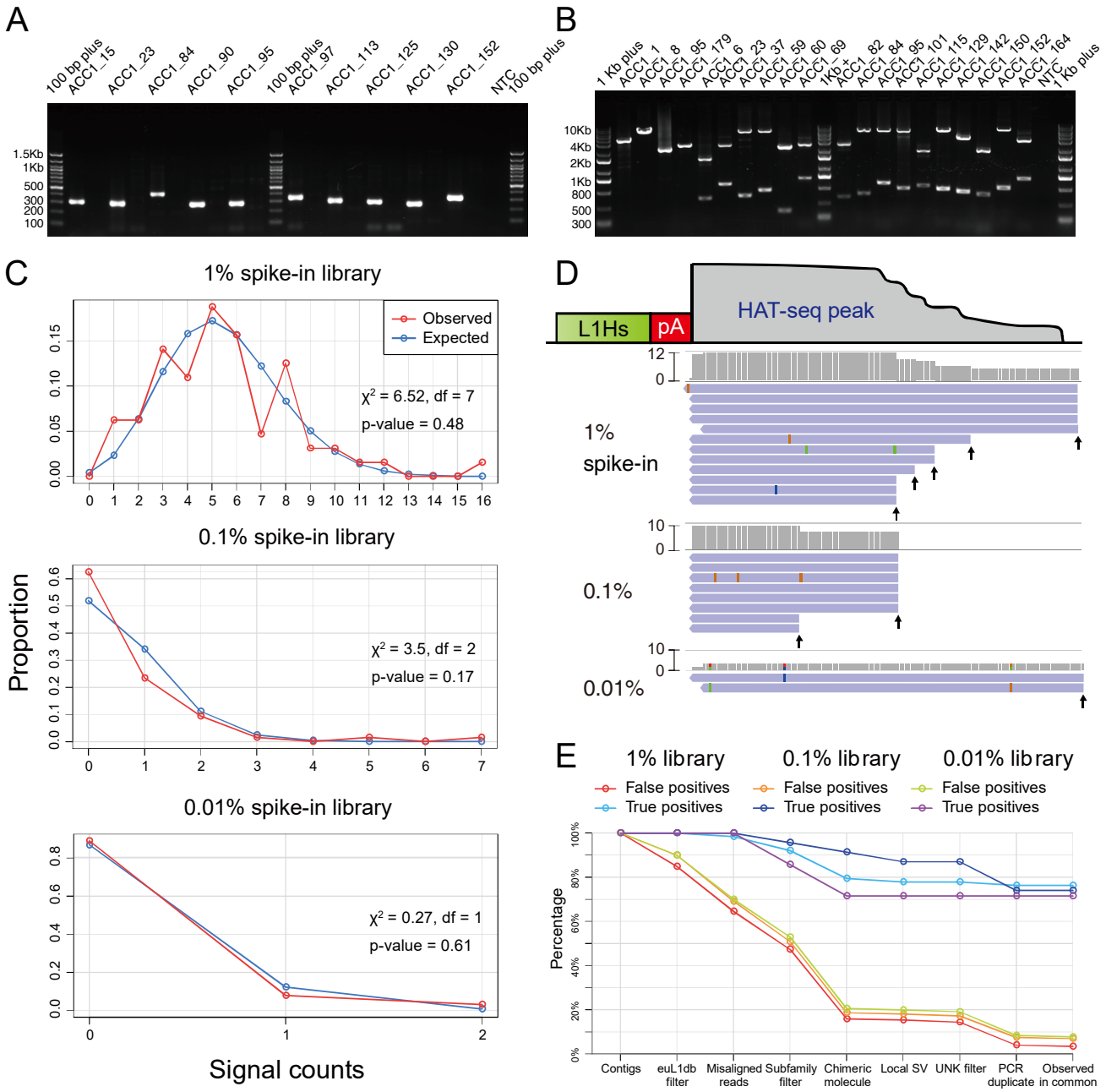


600

601

602 **Fig 1. Overview of human active transposon sequencing (HAT-seq).**

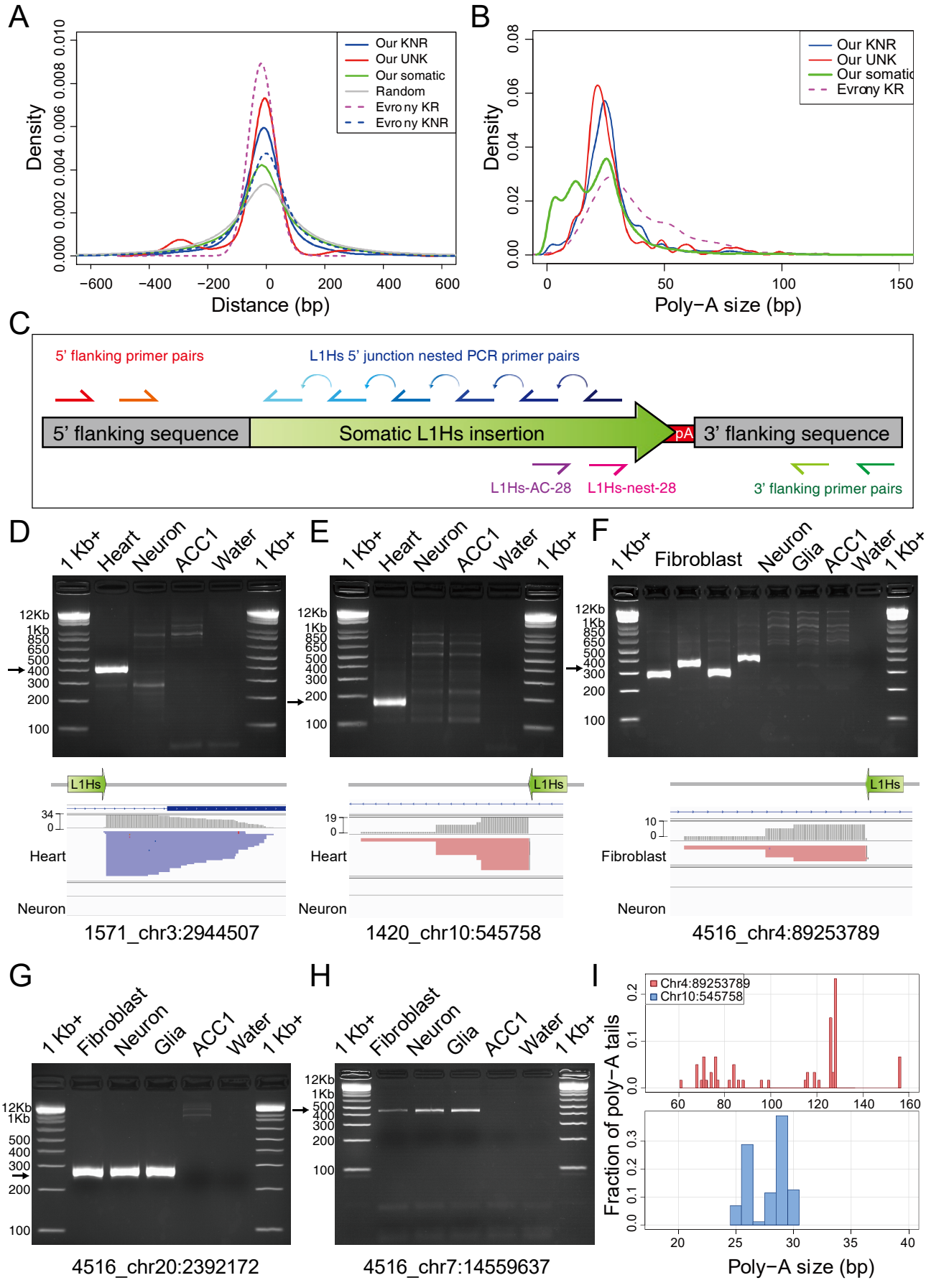
603 (A) Fluorescence-activated cell sorting (FACS) of prefrontal cortex (PFC) nuclei labeled with NeuN.
604 Two populations ($NeuN^+$ and $NeuN^-$) were sorted. (B) Schematic of the nucleotides-shifting design of
605 the HAT-seq method. By adding two, four, or six random nucleotides upstream of L1Hs-specific primer
606 (L1Hs-AC-28), we transformed the library from a uniform phase-0 amplicon library to a mixed library
607 with phase-2, phase-4, and phase-6 amplicons, which remarkedly improved the base calling accuracy
608 in Read 2. (C) HAT-seq libraries were sequenced with paired-end 150-bp reads. After merging paired
609 reads into contigs that fully spanned the L1Hs-genome 3' junction, genomic locations of each L1Hs
610 insertion were determined by the alignments of their 3' flanking genomic sequences.
611



612
613

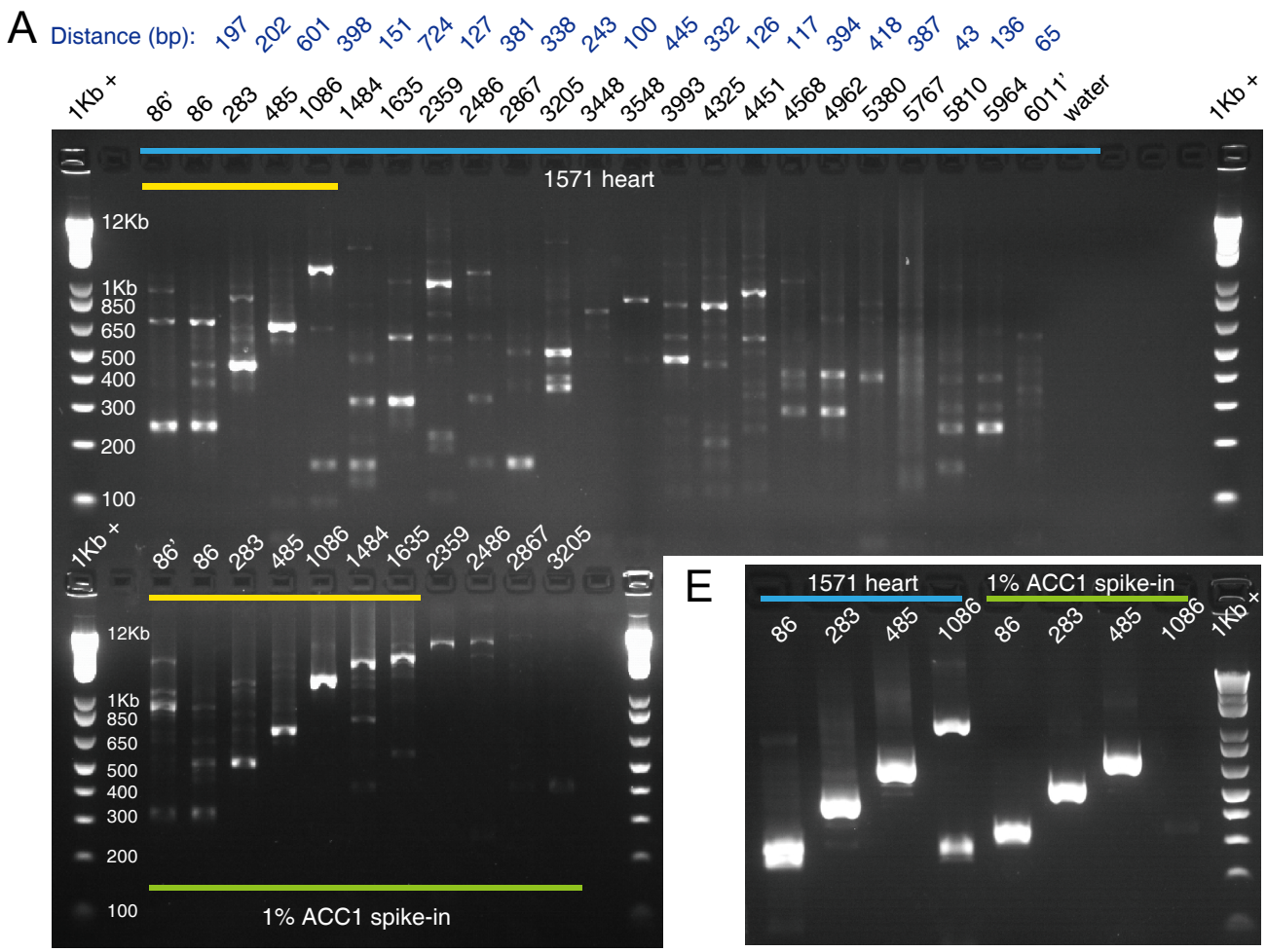
Fig 2. HAT-seq performance evaluation using a positive control.

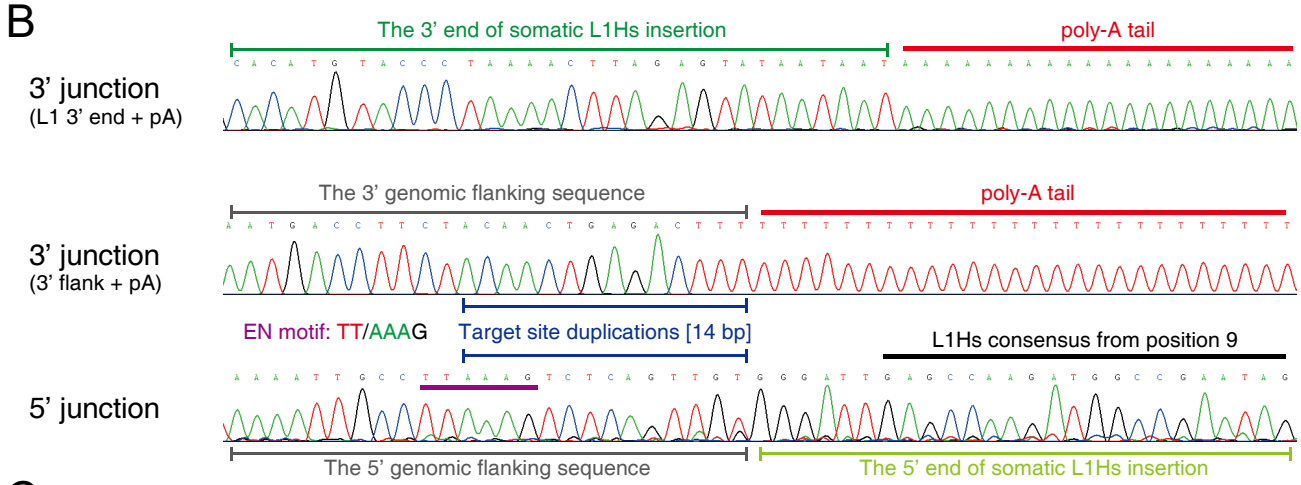
(A) Representative gel image used for the identification of ACC1-specific insertions based on 3' PCR analysis. For each site, genomic DNA from ACC1 and ACC2 was amplified using the same protocol. NTC: negative control. (B) Representative gel image used for the zygosity analysis of ACC1-specific insertions 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.



631 **Fig 3. Profiling of somatic L1Hs insertions in multiple human tissues.**

632 (A) The density distributions of L1 EN motifs around L1Hs integration sites. L1 EN motifs included
633 seven specific motifs (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG, CTAAAA, TCAAAA).
634 "Evrony KR" and "Evrony KNR" are germline L1Hs insertions identified in Evrony *et al.* 2012. (B)
635 The density distributions of poly-A tail length for each category of L1Hs insertion. (C) The PCR
636 validation scheme and locations of primers used. (D)–(H) Representative gel images of 3' nested PCR
637 validation for putative clonal somatic insertions. The Integrative Genomics Viewer screenshots for
638 (D)–(F) showed the coverage track (gray) and the alignment track (blue for read strand [-]; red for read
639 strand [+]) from HAT-seq data. Black arrows indicated bands with target size. 1Kb +: 1 Kb Plus DNA
640 ladder. (I) Polymorphic poly-A tail sizes of clonal somatic insertions. Top: fibroblast-specific somatic
641 L1Hs insertion at chr4:89253789 from Rett patient UMB#4516. Bottom: heart-specific somatic L1Hs
642 insertion at chr10:545758 from Rett patient UMB#1420.
643





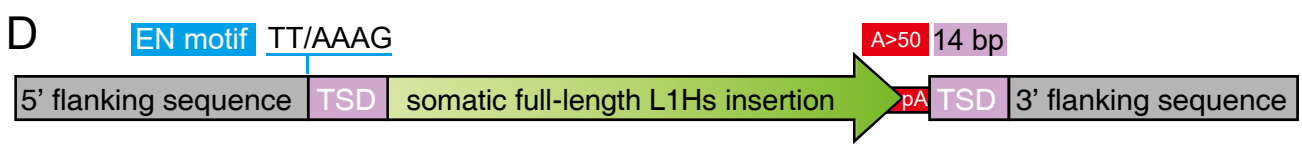
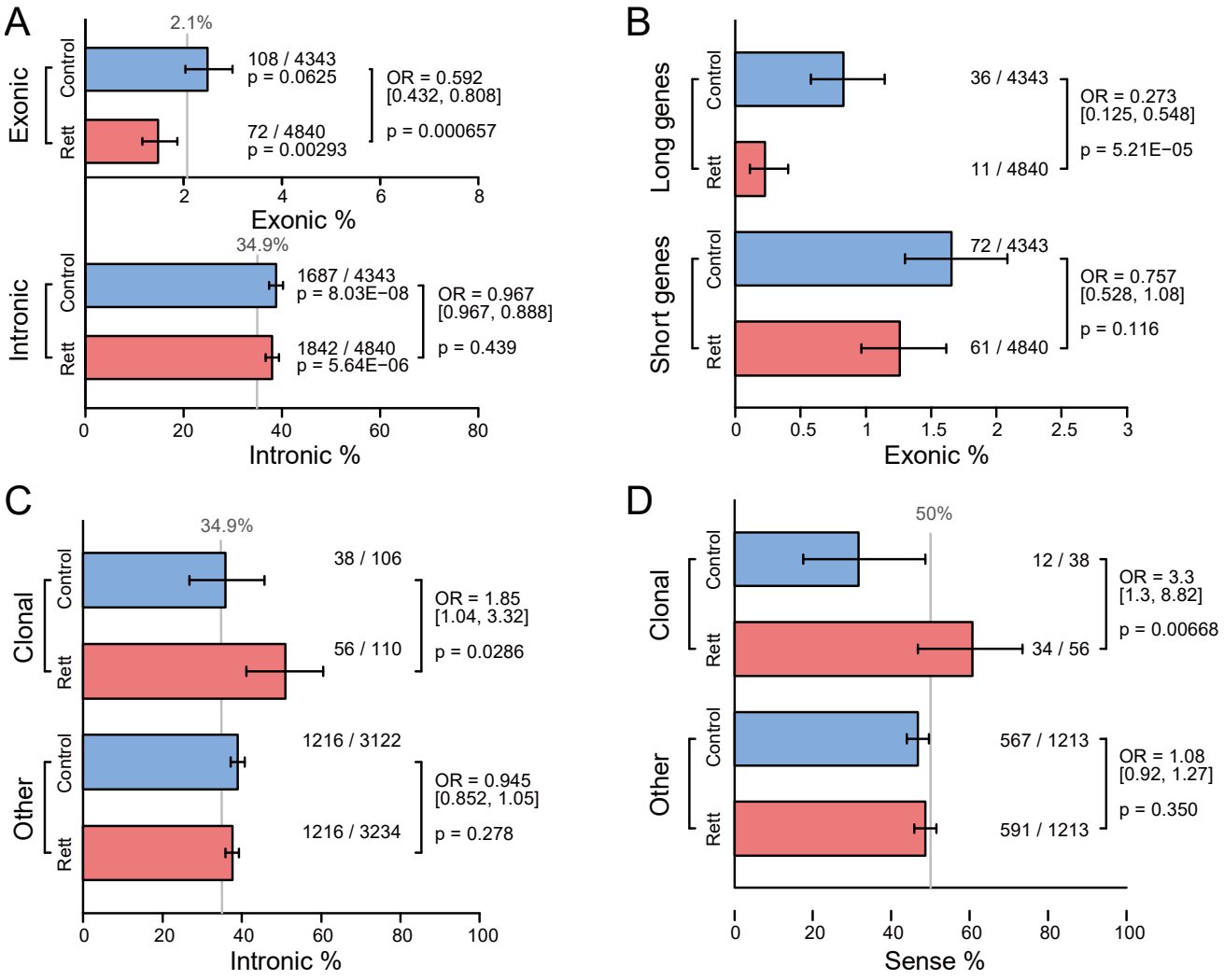


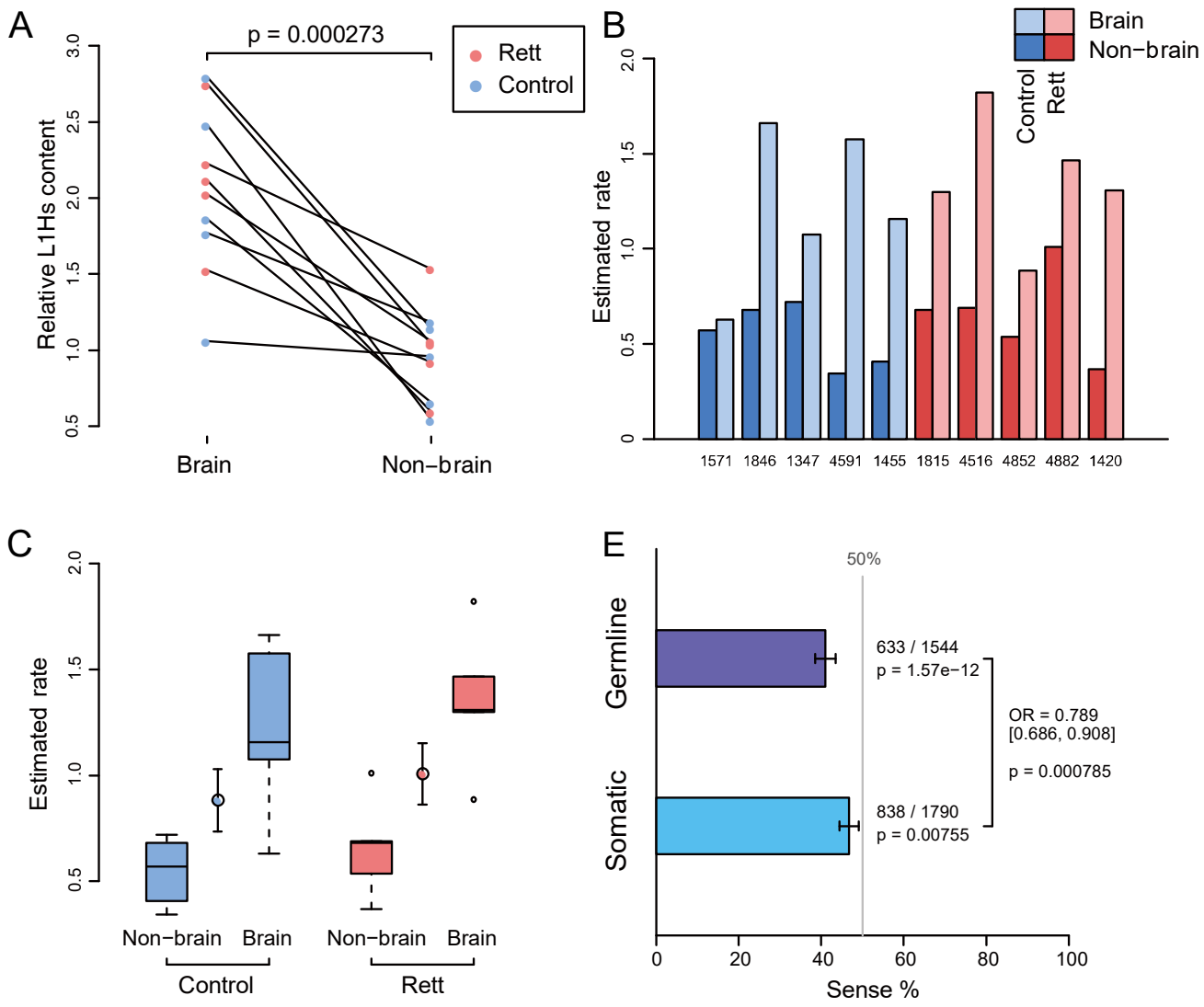
Fig 4. A full-length heart-specific L1Hs insertion (1571_chr3:2944507) in a healthy individual.

(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 Appendix 4) 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.



663 **Fig 5. Abnormal L1Hs mobilization in patients with Rett syndrome.**

664 (A) Percentages of somatic L1Hs insertions in exons and introns. (B) Percentages of somatic L1Hs
665 insertions in exons of long (> 100 kb) and short genes (< 100 kb). (C) Percentages of clonal somatic
666 L1Hs insertions in introns. (D) Percentages of sense-oriented clonal somatic L1Hs insertions. The gray
667 lines in (A) and (C) denoted the expected proportion determined by the exact base-pair count of that
668 specific region relative to the human genome. The gray line in (D) represented the expected proportion
669 if the insertions occurred randomly in both directions. Error bars in (A)–(D) indicated the 95%
670 confidence intervals.
671

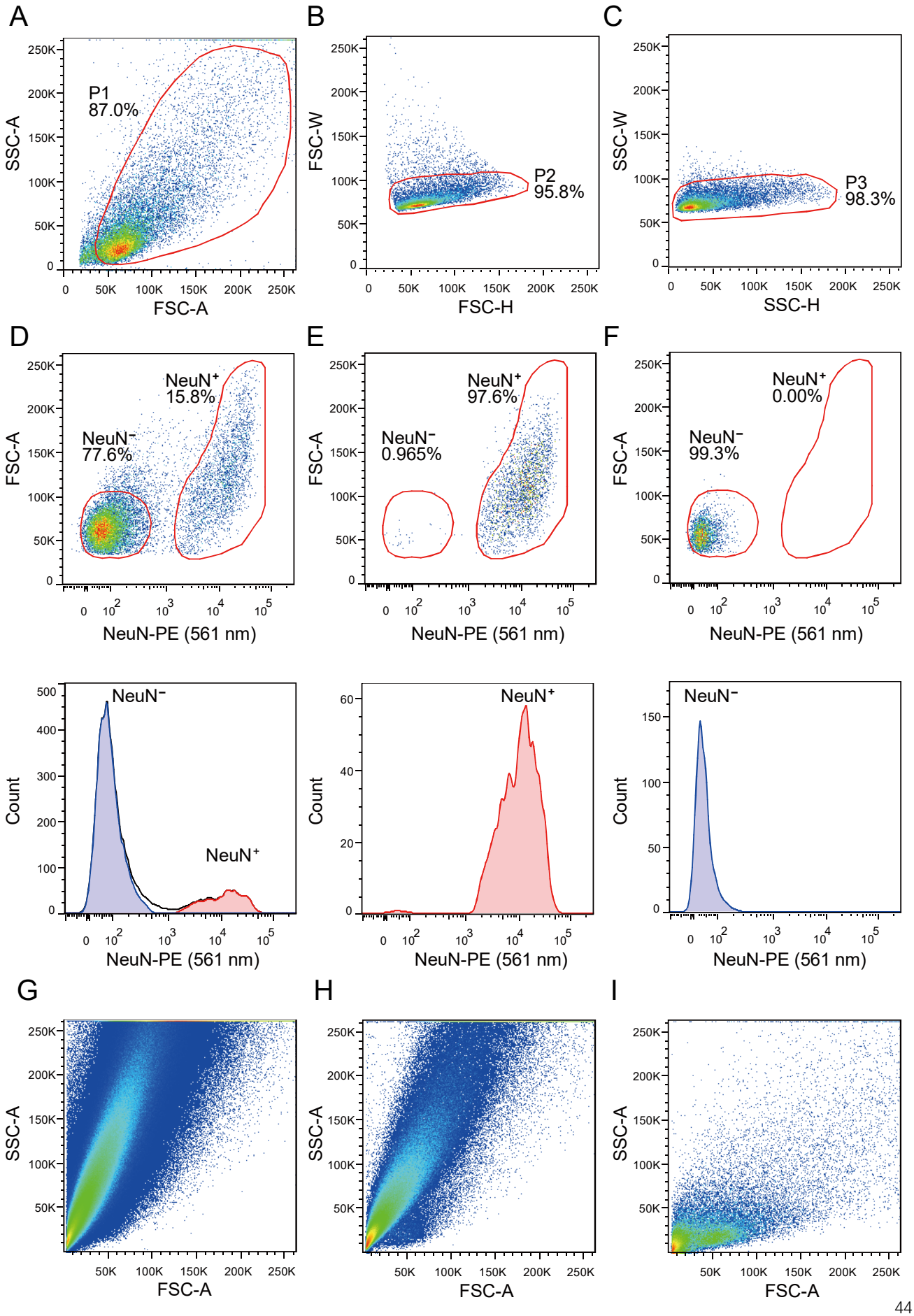


674 **Fig 6. Genome-wide patterns of somatic and germline L1Hs insertions.**

675 (A) Relative somatic L1Hs content in PFC neurons and non-brain tissue from the same donor. The
676 read count ratio of somatic insertions to germline KNR was calculated and then normalized relative to
677 the average value of non-brain samples. The linked dots represented pairs of brain and non-brain
678 samples obtained from the same individual. (B) Histogram of estimated rate of somatic L1Hs insertions
679 in each of tissue samples from the same donor based on the germline KNR copy number of each
680 individual. (C) Estimated rate of somatic L1Hs insertions for different tissue types and cohorts. Error
681 bars denoted the standard error of the mean (S.E.M). (D) Hierarchical clustering of all samples
682 sequenced in this study. Each row represented a sample, and each column represents an L1Hs germline
683 insertion. Black and white squares indicated the presence or absence of insertion, respectively. Column
684 annotations showed categories for known reference (KR; blue), known non-reference (KNR; green),
685 and unknown (UNK; red) insertions. (E) Percentages of sense-oriented germline and somatic L1Hs
686 insertions in transcripts. The gray line represented the expected proportion if the insertions occurred
687 randomly in both directions. Error bars indicated the 95% confidence intervals.
688

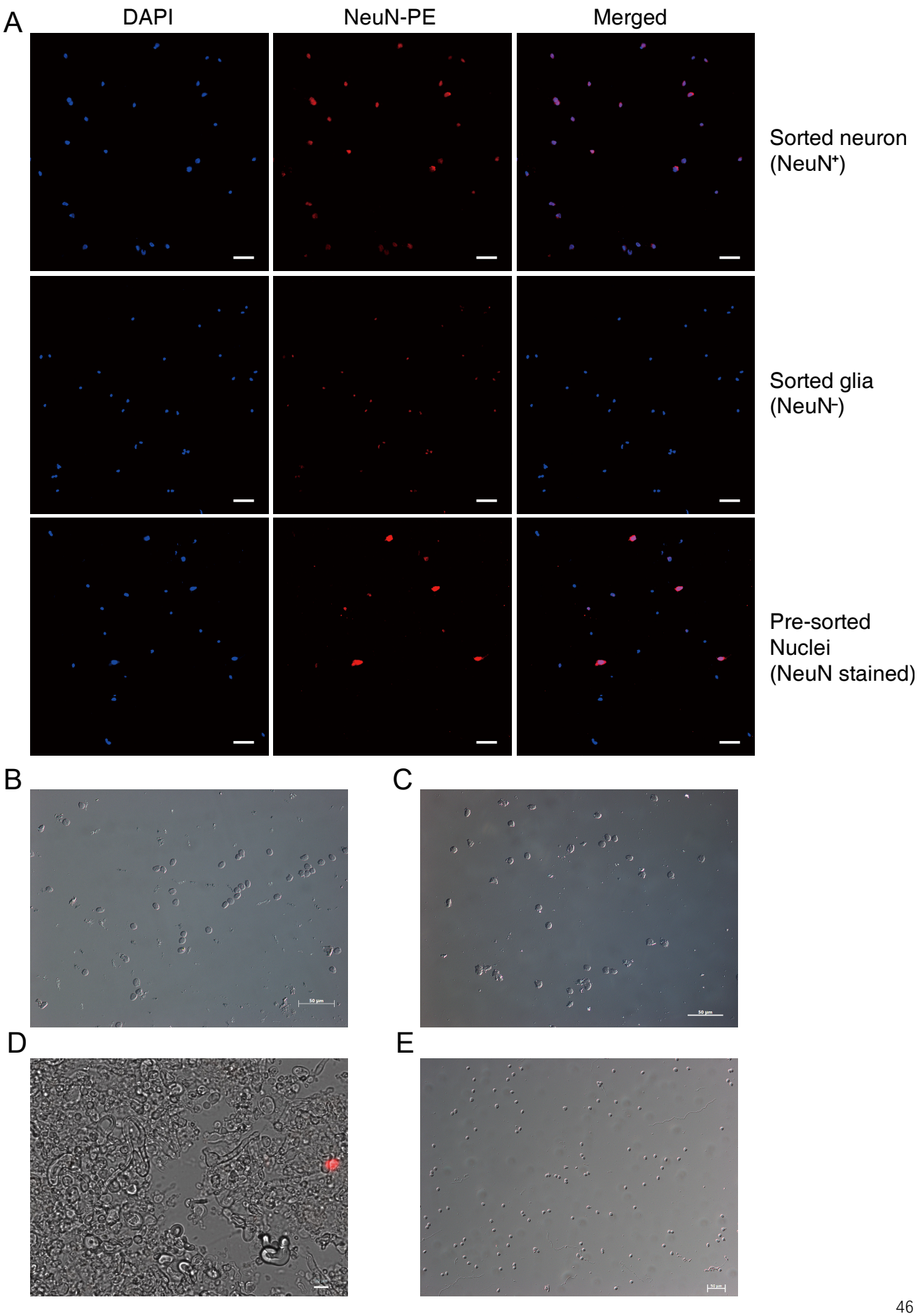
Supporting Information

Supplementary Figures



692 **S1 Fig. Nuclei isolation and NeuN$^+$ fluorescence-activated cell sorting (FACS).**

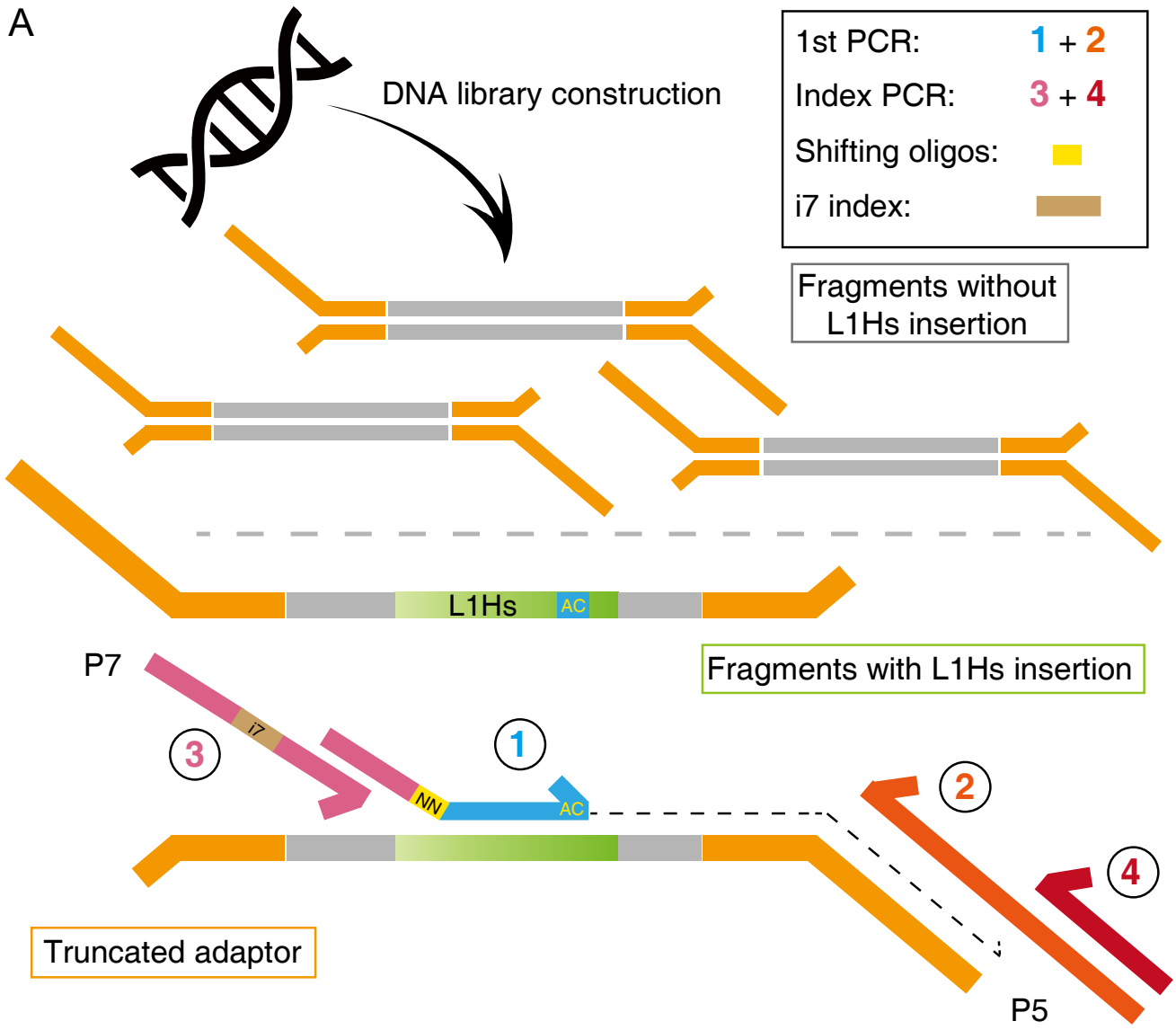
693 (A)–(D) Purify neuronal nuclei from human PFC. (A) The first gate (P1) was set as an FSC-A vs. SSC-
694 A plot to discriminate the population containing small-size debris. (B)–(C) The second (P2) and third
695 (P3) gates were set as FSC-H vs. FSC-W and SSC-H vs. SSC-D plots, respectively, to remove doublets
696 and clumps. (D) Top: NeuN$^-$ and NeuN$^+$ gates were set in the NeuN-PE (561 nm) vs. FSC-A plot.
697 Bottom: a count plot of NeuN-stained nuclei. (E) Purity analysis of sorted neurons. Top: sorted NeuN$^+$
698 nuclei were re-analyzed by FACS to confirm the sort purity. Bottom: a count plot of re-analyzed NeuN$^+$
699 nuclei. (F) Purity analysis of sorted glia. Top: sorted NeuN$^-$ nuclei were re-analyzed by FACS to
700 confirm the sort purity. Bottom: a count plot of re-analyzed NeuN$^-$ nuclei. (G) FSC vs. SSC plot of
701 brain homogenate. Brain homogenate contained a huge amount of cell debris and myelin debris. (H)
702 FSC vs. SSC plot of debris-detached single-nuclei homogenate. Minced brain tissue was soaked
703 overnight before homogenization and then incubated with nonionic detergent, Nonidet P-40, to remove
704 cell debris from nuclear membrane. (I) FSC vs. SSC plot of debris removed nuclei fraction. Cell debris
705 and myelin were separated from nuclei using Percoll density gradient centrifugation.
706

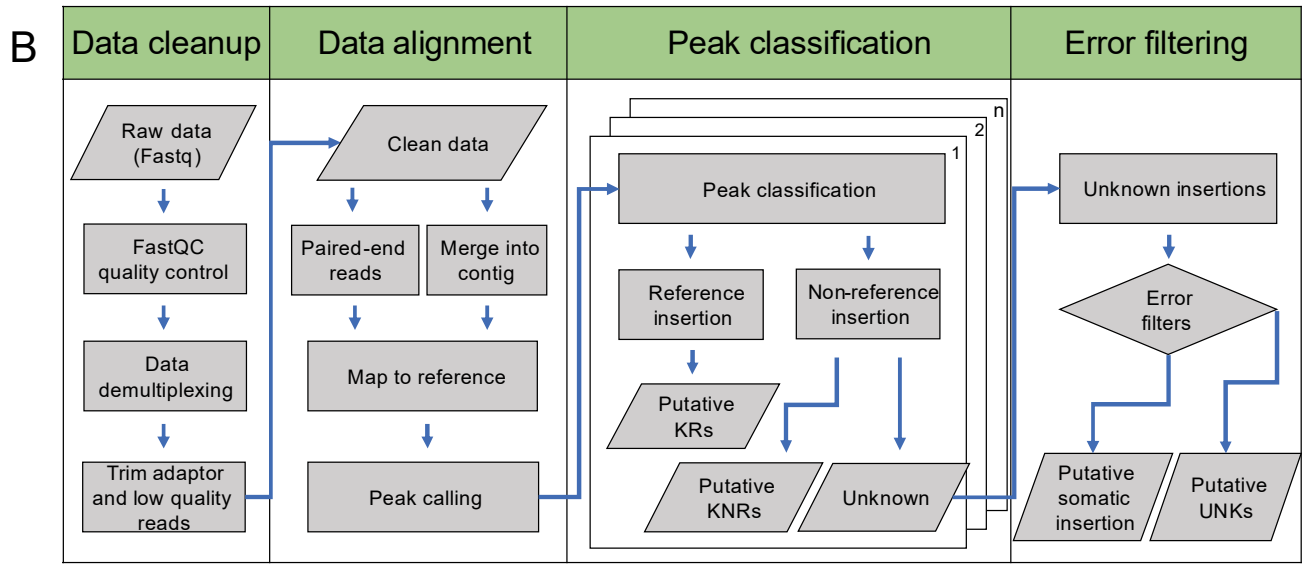


707

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.





720 **S3 Fig. Schematic diagrams of HAT-seq library construction and computational analysis pipeline.**

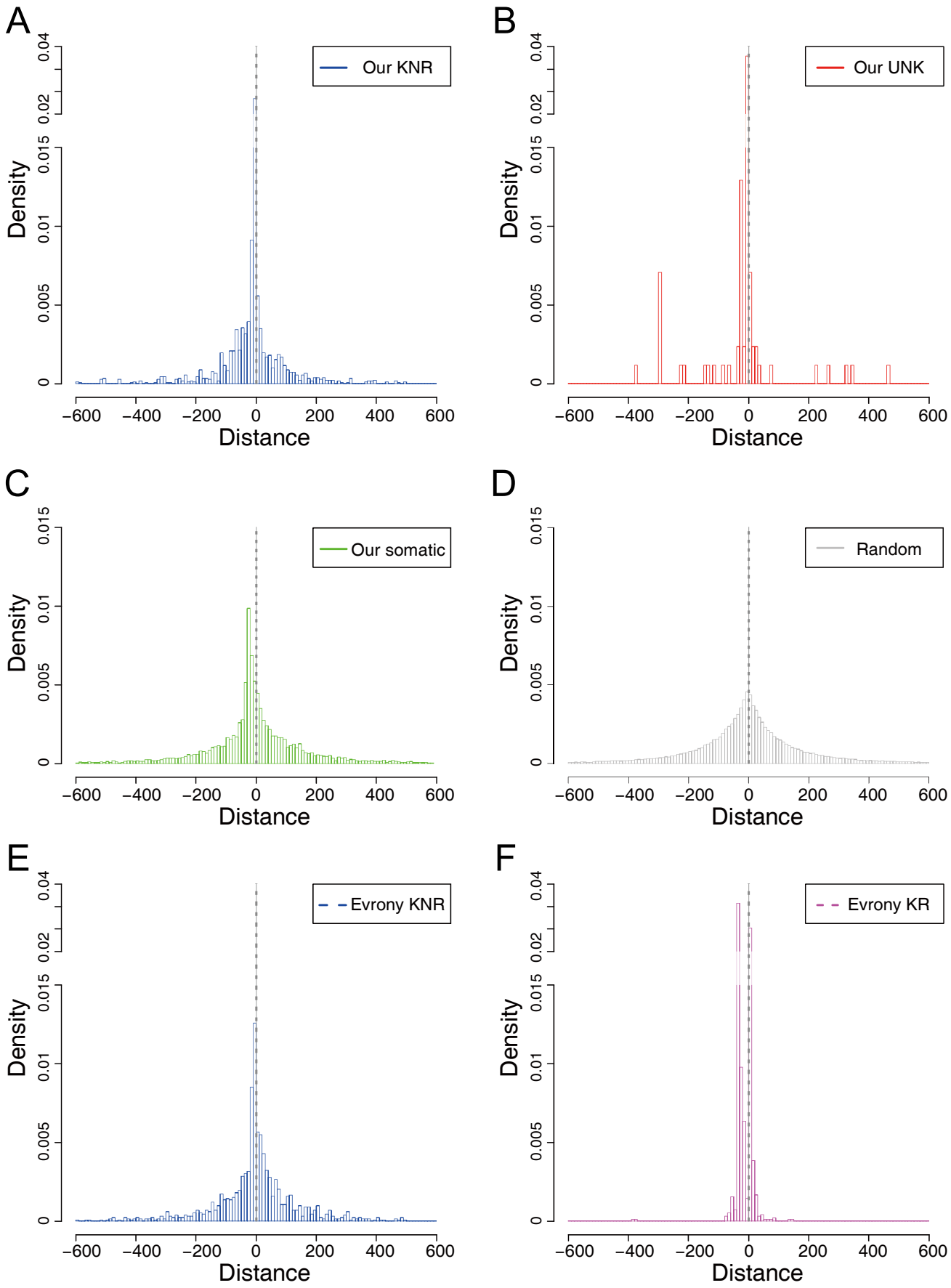
721 (A) Schematic of the HAT-seq library construction. The fragmented genomic DNA was ligated with

722 P7 truncated adaptors, and then used as template for L1Hs amplification PCR. Primers 1 (P7_Ns_L1Hs)

723 was specific to L1Hs diagnostic "AC" motif. See S1 Table for primer sequences. (B) Schematic of the

724 HAT-seq data analysis pipeline; full details are provided in the Materials and Methods.

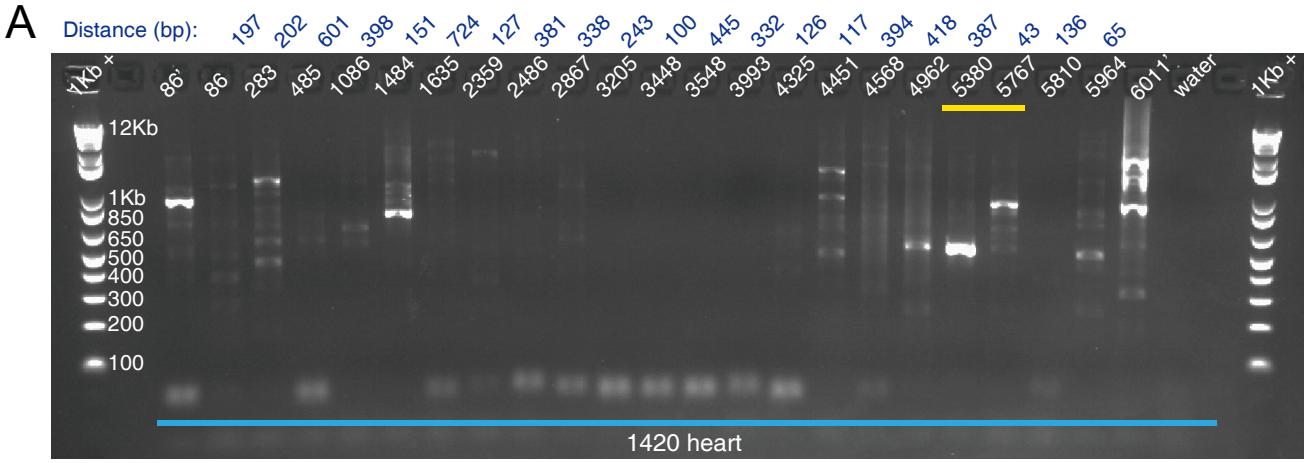
725

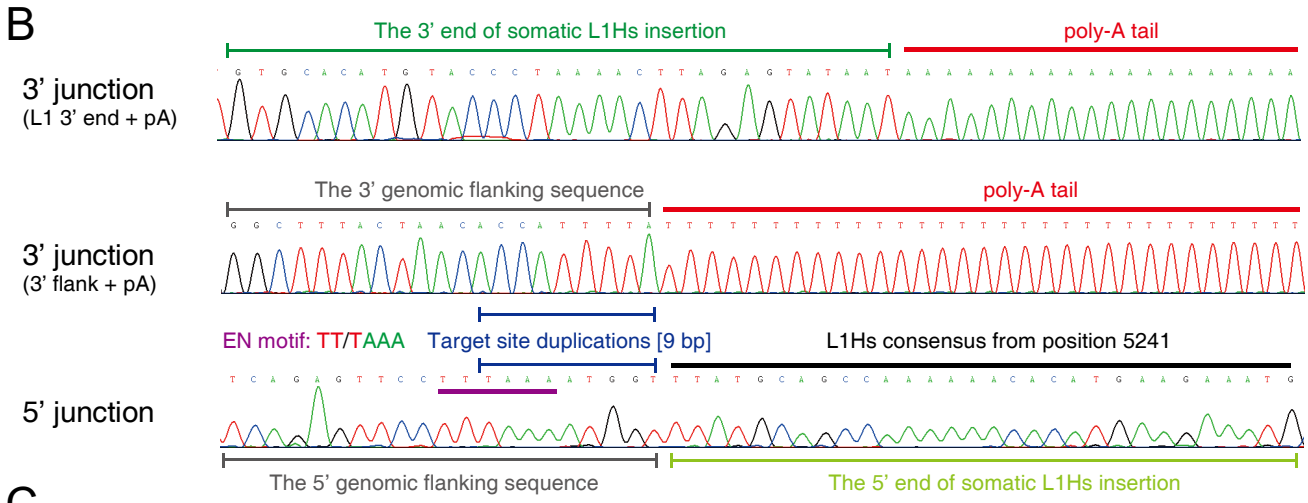


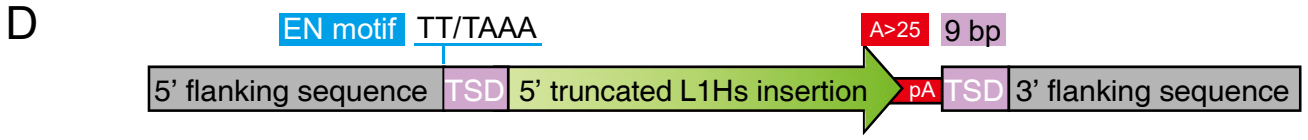
726
727

728 **S4 Fig. EN motif enrichment analysis across all categories of L1Hs insertions.**

729 The density distributions of L1 EN motifs around germline KNR (A), UNK (B), somatic insertions
730 (C), randomly sampled positions (D), “Evrony KR” (E), and “Evrony KNR” (F). The lists of “Evrony
731 KR” and “Evrony KNR” were extracted from Evrony *et al.* 2012. The bin size of histogram was 10bp.
732 L1 EN motifs included seven specific motifs (TTAAAA, TTAAGA, TTAGAA, TTGAAA, TTAAAG,
733 CTAAAA, TCAAAA).
734

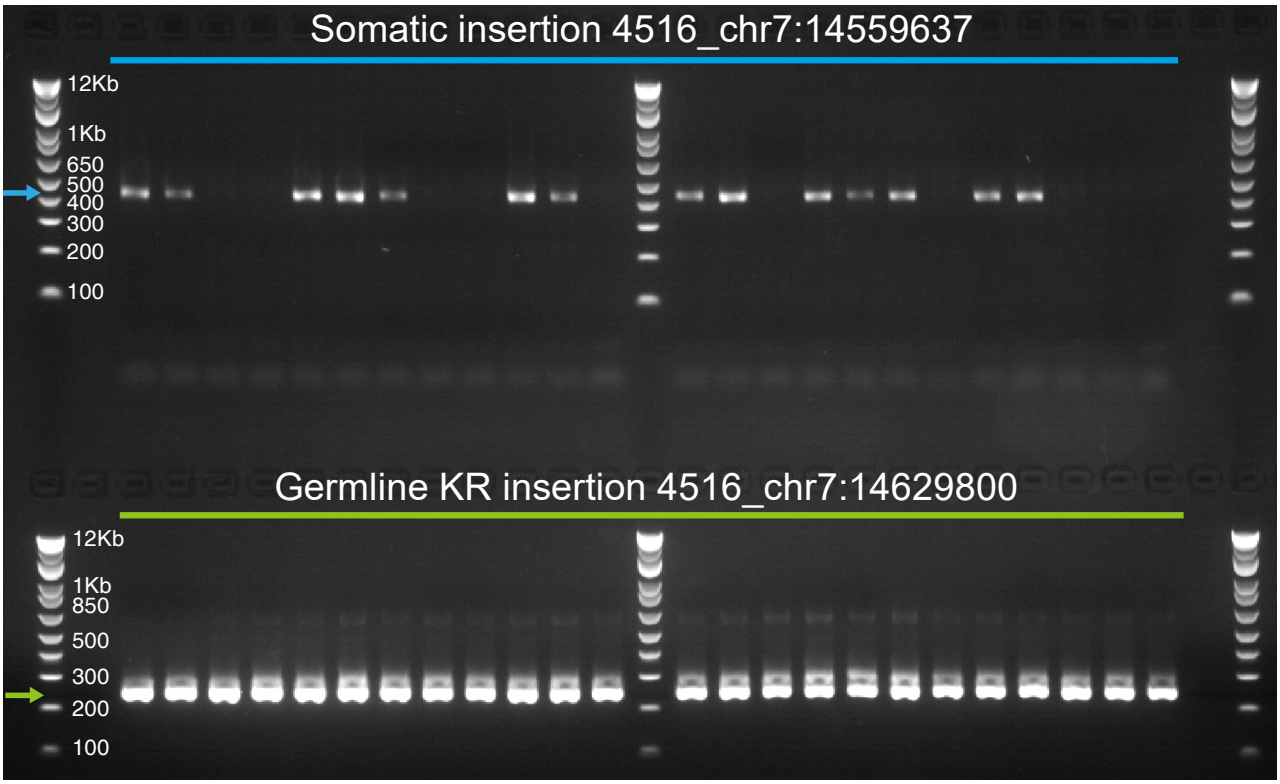






737 **S5 Fig. A 5' truncated heart-specific L1Hs insertion (1420_chr10:545758) in a Rett patient.**

738 (A) The agarose gel image of 5' junction nested PCR validation for the heart-specific L1Hs insertion
739 in the Rett patient (UMB#1420). The locations of primers used in 5' junction PCR assays were labeled
740 on the top of each lane, where primers with the prime symbol denoted semi-nested PCR assays. The
741 distances between each two adjacent 5' step-wise primers were labeled on the top (dark blue). The
742 yellow line highlighted the expected stair-step bands in 5' junction PCR. 1Kb +: 1 Kb Plus DNA ladder.
743 (B) The Sanger sequencing chromatograms of the 3' and 5' junctions of the somatic insertion
744 (1420_chr10:545758). The L1 EN motif and TSD were indicated by purple and blue lines. (C) Multiple
745 sequence alignment of the 5' end between the identified somatic insertion and three L1Hs consensus
746 sequences (L1Hs Repbase consensus and two hot L1s in human [L1.3 and L1.4]). (D) The schematic
747 structure of the highly 5' truncated (~800 bp) L1Hs insertion 1420_chr10:545758.
748



749
750

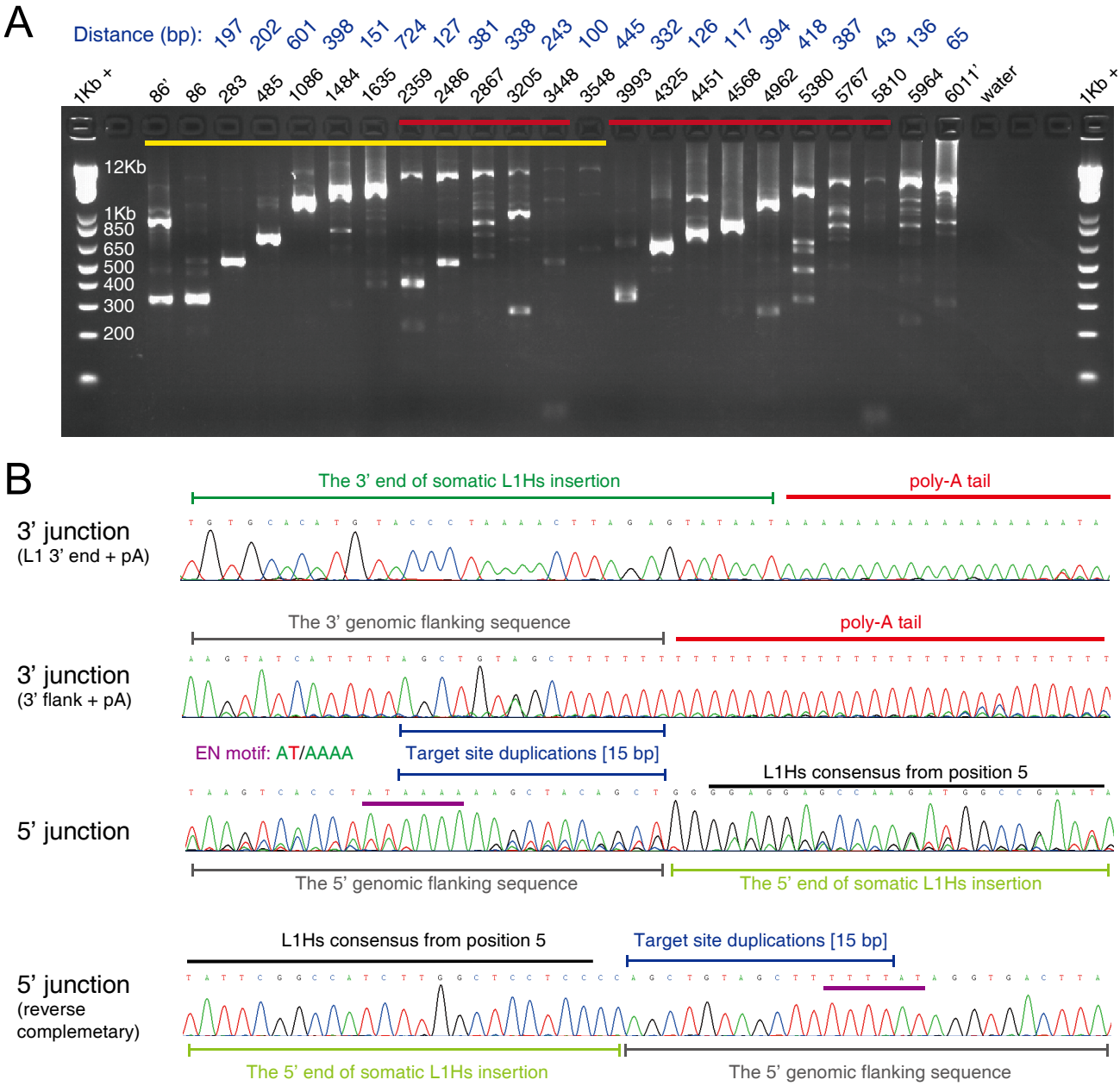
751 **S6 Fig. The somatic status of an L1Hs insertions (4516_chr7:14559637) in a Rett patient.**

752 The somatic insertion 4516_chr7:14559637 was present in 14 out of 24 nested 3' PCR wells, compared

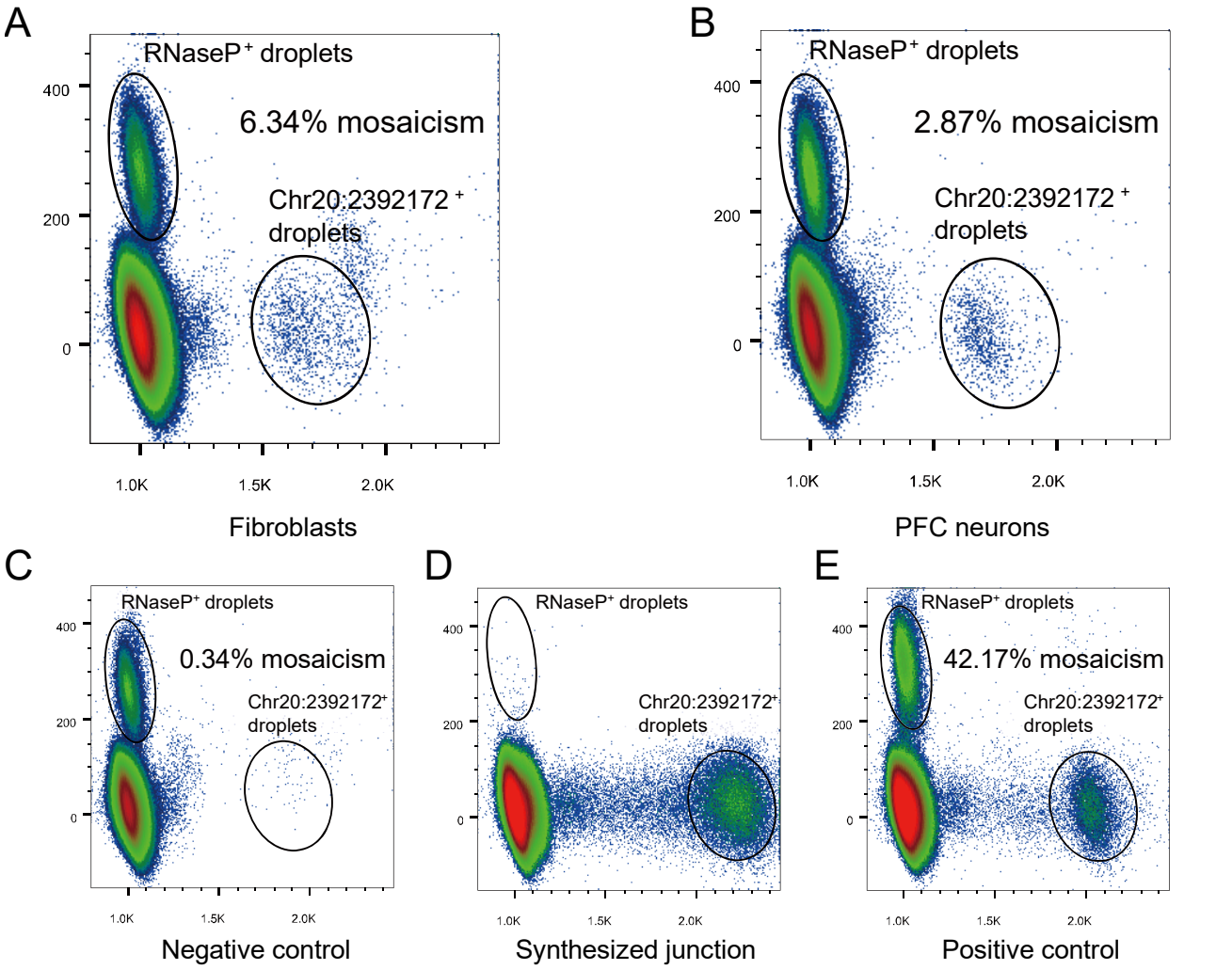
753 to 24 out of 24 wells for a germline KR insertion (chr7:14629800) from the same donor. DNA sample

754 was diluted to ~300 cells per well. Blue and green arrows indicated bands with target size.

755



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_chr20:2392172). 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.



775 **S8 Fig. Qualitative and quantitative analysis of L1Hs insertions.**

776 (A)–(E) Droplet digital PCR (ddPCR) assays to quantify mosaicism (percentage of cells) of somatic
777 L1Hs insertions at chr20:2392172 in fibroblasts (A) and PFC neurons (B) from Rett patient
778 UMB#4516. Fragmented ACC1 blood gDNA was used as template for negative control assay (C). A
779 mixed template containing fragmented ACC1 blood gDNA and diluted synthesized L1Hs genome
780 junction oligos (D) was used for positive control assay (E). RNaseP served as a genomic copy number
781 reference (copy number = 2). L1Hs and RNaseP assays were labeled with FAM and VIC, respectively.
782 (F)–(G) Relative somatic L1Hs content in PFC neurons and non-brain tissue from the same donor,
783 normalized by the read count of KRs (F) or UNKs (G) from the same tissue sample.
784

785 **Supplementary Tables**

786

787 **S1 Table. Primer sequences of HAT-seq library**

788

789 **S2 Table. ACC1-specific insertions in positive control experiments**

790

791 **S3 Table. Statistics of error filters in positive control experiments**

792

793 **S4 Table. Clinical characterization of patients with Rett syndrome**

794

795 **S5 Table. Statistics of HAT-seq libraries**

796

797 **S6 Table. Statistics of known reference insertions among all samples**

798

799 **S7 Table. Statistics of polymorphic insertions among all samples**

800

801 **S8 Table. Statistics of somatic insertions among all samples**

802

803 **S9 Table. TPRT hallmark annotation for all somatic insertions**

804

805 **S10 Table. 3' junction nested PCR and digital droplet PCR validation**

806

807 **S11 Table. 5' junction nested PCR validation**

808

809 **S12 Table. Annotation for somatic exonic insertions**

810

811 **S13 Table. Raw data for statistical analyses**

812

813 **S14 Table. Quantification statistics among all samples**

814

815 **S15 Table. L1Hs enrichment analysis on HAT-seq**

816

817 **S16 Table. 3' junction PCR validation for germline insertions**

818

819 **Supplementary Files**

820

821 **Appendix 1. Cell type-specific sorting for postmortem human brain samples**

822

823 **Appendix 2. Identification of ACC1-specific insertions and their zygosity**

824

825 **Appendix 3. Detection of somatic insertions in positive control experiments**

826

827 **Appendix 4. Benchmarking PCR validation assays for low-frequency somatic insertions**

828

829 **Appendix 5. Experimental validation of polymorphic germline L1Hs insertions**