Cell Research (2018) 28:187-203.

www.nature.com/cr

ORIGINAL ARTICLE

Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders

Jasmine Jacob-Hirsch[1, 2], Eran Eyal[1], Binyamin A Knisbacher[2], Jonathan Roth[3, 5], Karen Cesarkas[1], Chen Dor[1], Sarit Farage-Barhom[1], Vered Kunik[1], Amos J Simon[1], Moran Gal[2], Michal Yalon[4], Sharon Moshitch-Moshkovitz[1], Rick Tearle[6], Shlomi Constantini[3, 5], Erez Y Levanon[2], Ninette Amariglio[1, 2], Gideon Rechavi[1, 5]

[1]Cancer Research Center and the Wohl Institute of Translational Medicine, the Chaim Sheba Medical Center, Tel Hashomer, Israel; [2]Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Israel; [3]Department of Pediatric Neurosurgery, Dana Children's Hospital, Tel Aviv Medical Center, Israel; [4]Department of Pediatric Hematology-Oncology, Edmond and Lily Safra Children's Hospital, The Chaim Sheba Medical Center, Tel Hashomer, Israel; [5]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel; [6]Complete Genomics, 2071 Stierlin Court, Mountain View, CA 94043, USA

Neural progenitor cells undergo somatic retrotransposition events, mainly involving L1 elements, which can be potentially deleterious. Here, we analyze the whole genomes of 20 brain samples and 80 non-brain samples, and characterized the retrotransposition landscape of patients affected by a variety of neurodevelopmental disorders including Rett syndrome, tuberous sclerosis, ataxia-telangiectasia and autism. We report that the number of retrotranspositions in brain tissues is higher than that observed in non-brain samples and even higher in pathologic vs normal brains. The majority of somatic brain retrotransposons integrate into pre-existing repetitive elements, preferentially A/T rich L1 sequences, resulting in nested insertions. Our findings document the fingerprints of encoded endonuclease independent mechanisms in the majority of L1 brain insertion events. The insertions are "non-classical" in that they are truncated at both ends, integrate in the same orientation as the host element, and their target sequences are enriched with a CCATT motif in contrast to the classical endonuclease motif of most other retrotranspositions. We show that L1Hs elements integrate preferentially into genes associated with neural functions and diseases. We propose that pre-existing retrotransposons act as "lightning rods" for novel insertions, which may give fine modulation of gene expression while safeguarding from deleterious events. Overwhelmingly uncontrolled retrotransposition may breach this safeguard mechanism and increase the risk of harmful mutagenesis in neurodevelopmental disorders.

***Keywords*:** neurodevelopmental disorders; brain; retrotransposition; L1Hs; Rett syndrome; tuberous sclerosis
Cell Research (2018) **28**:187-203. doi:10.1038/cr.2018.8; published online 12 January 2018

Introduction

Half of the human genome is derived from transposable genetic elements, mainly retrotransposons that self-replicate through reverse transcription of RNA intermediates. These retrotransposons include long terminal repeat (LTR) elements as well as non-LTR long and short interspersed elements (LINEs and SINEs, respectively) [1-3]. Some members of three non-LTR retrotransposon families, namely L1 (LINE1), Alu and SVA have been active in recent human genome evolution [3, 4]. The mutagenic activity of retrotransposons can be detrimental, and so far around 100 disease cases resulting from heritable or *de novo* retrotransposition events have been described [4]. The most flourishing of all human mobile elements are L1 and Alu; the former comprise 17% and the latter, 11% of the human genome. The full length L1 element is 6 kb long and capable of autonomous retrotransposition, mainly via target-primed reverse transcription (TPRT) [5]. About 900 000 copies of L1-derived sequences, including numerous fragmented elements, most of them transposition incompetent, are annotated in the human genome (hg19, RepeatMasker; RM, URLs section). L1Hs, the youngest human-specific L1 subfamily, encompasses a minor subset of "hot" elements

Correspondence: Gideon Rechavi
E-mail: gidi.rechavi@sheba.health.gov.il
Received 17 July 2017; revised 10 November 2017; accepted 20 November 2017; published online 12 January 2018

actively involved in most current insertions [6]. The 0.3 kb long, primate specific, Alu elements have a copy number well in excess of one million [1, 2, 4, 5]; they are non-autonomous and depend on L1 trans-acting factors for their retrotransposition. The youngest and most active Alu subfamily is AluY, mainly its AluYa5 and AluYb8 members [1, 4, 7]. Although transposition of mobile genetic elements in metazoans was long considered a germline-specific phenomenon, it was demonstrated three decades ago that insertions of such elements occur also in somatic cells, particularly in cancer [8-10]. Whereas in the past, identification of such rare retrotransposition events was serendipitous, the advance of next generation sequencing (NGS) has enabled an active prospective search for transposition events based on the analysis of captured elements or of whole genomes. NGS-based approaches have allowed the detection of retrotransposition events in various types of tumors [11, 12]. Somatic L1 retrotranspositions were reported not only in cancer but, interestingly, also in rodent and human neural progenitor cells (NPCs). These insertion events contribute to somatic mosaicism and to increased complexity in the brain [13-18]. The fact that retrotransposition occurs normally during neural differentiation raises the question how such a stochastic mechanism, that may lead to hazardous mutagenic events, was not selected against. Indeed, enhanced brain L1 retrotransposition was demonstrated in patients affected by two genetic disorders, Rett syndrome (Rett) [19] and ataxia-telangiectasia (AT), illustrating the potential risk of enhanced retrotransposition [20]. Rett is a neurodevelopmental disorder resulting from a mutation in the X-linked *MECP2* gene [19] and is characterized by slow development, seizures, intellectual disability and autistic-like behaviors [21]. AT is an inherited disorder resulting from mutations in the *ATM* gene, which affect primarily the nervous and immune systems [22]. The identification of increased L1 retrotransposition in Rett and AT was based on the study of NPCs using L1-EGFP reporter and on quantitative PCR that showed increased L1 DNA copy number [19, 20]. However, no whole-genome sequencing (WGS) of Rett or AT brain tissues was reported so far. Here we quantified, characterized and confirmed, using WGS, the increased retrotransposition in brain samples from individuals with these syndromes. In addition, we analyzed by WGS two additional neurodevelopmental disorders, tuberous sclerosis complex (TSC) and non-syndromic autism (NSA) where no information regarding brain retrotransposition is available. TSC is a multi-system genetic disease, resulting from mutations in the *TSC1* or *TSC2* genes [23]. TSC patients are commonly affected by seizures, developmental delay, intellectual disability and autistic behaviors as well as by tumors, including distinctive brain tumors known as subependymal giant cell astrocytomas (SEGA) that are derived from NPCs [24, 25]. NSA defines cases where autism is the primary diagnosis and not secondary to a disorder caused by a well-known mutated gene, such as *Rett* and *TSC* [26]. We show here increased retrotransposition in TSC, in particular in the TSC-related SEGA tumors, as well as in NSA. Our findings therefore expand the spectrum of neurodevelopmental disorders in which enhanced retrotransposition has been demonstrated, and provide, for the first time, comprehensive whole-genome mapping of brain insertional mutagenesis. We describe here unique and novel features of retrotransposition in the brain, that allow the "safe landing" of most L1 insertions. Such L1 transpositions are assumed to increase neuronal diversity and plasticity [14-16, 27] while minimizing the risk of deleterious mutagenic effects. Overwhelmingly uncontrolled retrotransposition that breaches the insertion safeguard mechanisms increases the risk of harmful mutagenesis and may contribute to neuropathology.

Results

Prevalent novel mobile element insertions in brain tissues

To study the brain retrotransposition phenomenon in depth, we sequenced 20 normal and pathological brain samples by the Complete Genomics (CG) technology (Supplementary information, Figure S1A-S1F). The WGS was performed with a minimal depth of 40× and an average depth of 80×, allowing efficient mapping of repetitive elements and detection of somatic events. We sequenced whole genomes of human normal brain (NB) and brain tissues of patients affected by the neurodevelopment disorders: Rett, AT, NSA, TSC including the unique TSC-archetypal SEGA tumors with their paired peripheral blood (PB) samples (Table 1). To explore the novel retrotransposon landscape and as a reference for comparison with the genomes of the various brain samples, we analyzed whole genomes of 78 lymphoblastoid cell lines derived from normal individuals representing 11 different human populations (LB1-78; "CG Public Genome Data Repository", URLs section, Table 1, Supplementary information, Table S1A and S1B). Using a stringent approach (detailed in Supplementary information, Figure S2) we identified 111 671 novel AluY insertion events and 123 862 novel L1Hs insertion events in the 100 samples analyzed, including both somatic and germline polymorphic events that are not included in the RM annotation (Figure 1A, 1B, Table 1 and Supplementary information, Table S1A). The numbers of novel

Table 1 Number of novel mobile element insertions (MEI)

Sample group	Sample name	Donor	Total AluY insertion count[a]	Total L1Hs insertion count[a]	Tissue origin
Ataxia-telangiectasia (AT)	AT1[b]	1	1 637	975	Cerebellum
	AT2[b]	1	1 589	27 692	Frontal cortex
	AT3	2	1 343	941	Frontal cortex
	AT4	3	1 392	34 523	Frontal cortex
Non syndromic autism (NSA)	NSA1[b]	4	1 817	2 162	Cerebellum
	NSA2[b]	4	1 701	1 137	Frontal cortex
Rett syndrome	Rett 1[b]	5	1 422	2 233	Cerebellum
	Rett 2[b]	5	1 386	2 274	Frontal cortex
Subependymal giant cell	SEGA1	6	1 485	11 310	SEGA
Astrocytoma (SEGA)	SEGA2	7	1 509	1 669	SEGA
	SEGA3[b,c]	8	1 485	9 955	SEGA
SEGA patients peripheral blood (PB)	PB1[b]	6	1 411	317	Peripheral blood
	PB3[b]	8	1 437	416	Peripheral blood
Tuberous sclerosis (TSC)	TSC1[b]	9	1 375	630	Cerebellum
non-tumoral brain tissues	TSC2[b]	9	1 357	911	Occipital cortex
	TSC3[b]	10	1 419	937	Cerebellum
	TSC4[b]	10	1 444	1 506	Occipital cortex
Normal brain (NB)	NB1	11	1 333	625	Hippocampus
	NB2[b]	12	1 336	943	Cerebellum
	NB3[b]	12	1 319	1 127	Occipital cortex
	NB4	13	1 406	591	Cerebellum
	NB5	14	1 411	582	Cerebellum
Normal lymphoblastoid (LB) cell lines[c]	Normal LB[d]	15-92	1 021.2	261.6	Lymphoblastoid cell lines

[a]Read number ≥ 10, insertion length > 45 bases, insertion score > 50
[b]Samples are from the same donor; see "Donor" column
[c]SEGA3 was surgically excised from a patient with no TSC syndrome
[d]Average of 78 LB samples

mobile element insertions (MEIs) (L1Hs and AluY) detected in brain tissues were higher than those in non-brain control samples ($P < 0.0001$; Mann-Whitney test). The number of L1Hs insertions was 8.5 fold higher in the pathological brain samples compared to the NB samples (average of 6 590.3 vs 773.6; $P < 0.05$; Mann-Whitney test) indicating that excessive L1Hs retrotransposition is a common characteristic of neurodevelopmental disorders. Figure 1C and 1D show the number of novel L1Hs and AluY events that are also found in the 1000 Genomes Project (1000G) MEI data, indicating that they are non-reference germline polymorphic events. The fact that the distribution of the number of polymorphic insertions is homogenous among all the samples, while the total number of L1Hs insertion events is skewed towards the brain samples, points to a high rate of somatic L1Hs insertion events in the brain samples (Figure 1E). The above findings together with the detection of most (84%) of the known polymorphic events reported in the 1000G MEI data [28] support the authenticity of MEI detection by the CG methodology (Supplementary information, Data S1). The accuracy of the methodology was also verified by PCR followed by Sanger sequencing as well as by L1 capturing followed by Illumina NGS (Supplementary information, Figure S3 and Tables S2, S3A, S3B). We also confirmed the unique characteristics of L1Hs insertions by analyzing the Pacific Biosciences data (Figure 2 and Supplementary information, Figure S4).

The mapping of an unprecedented number of MEIs by an unbiased WGS approach enabled novel key characteristics of brain retrotransposition to be deciphered and revealed that enhanced retrotransposition is a common feature of various neurodevelopmental disorders.

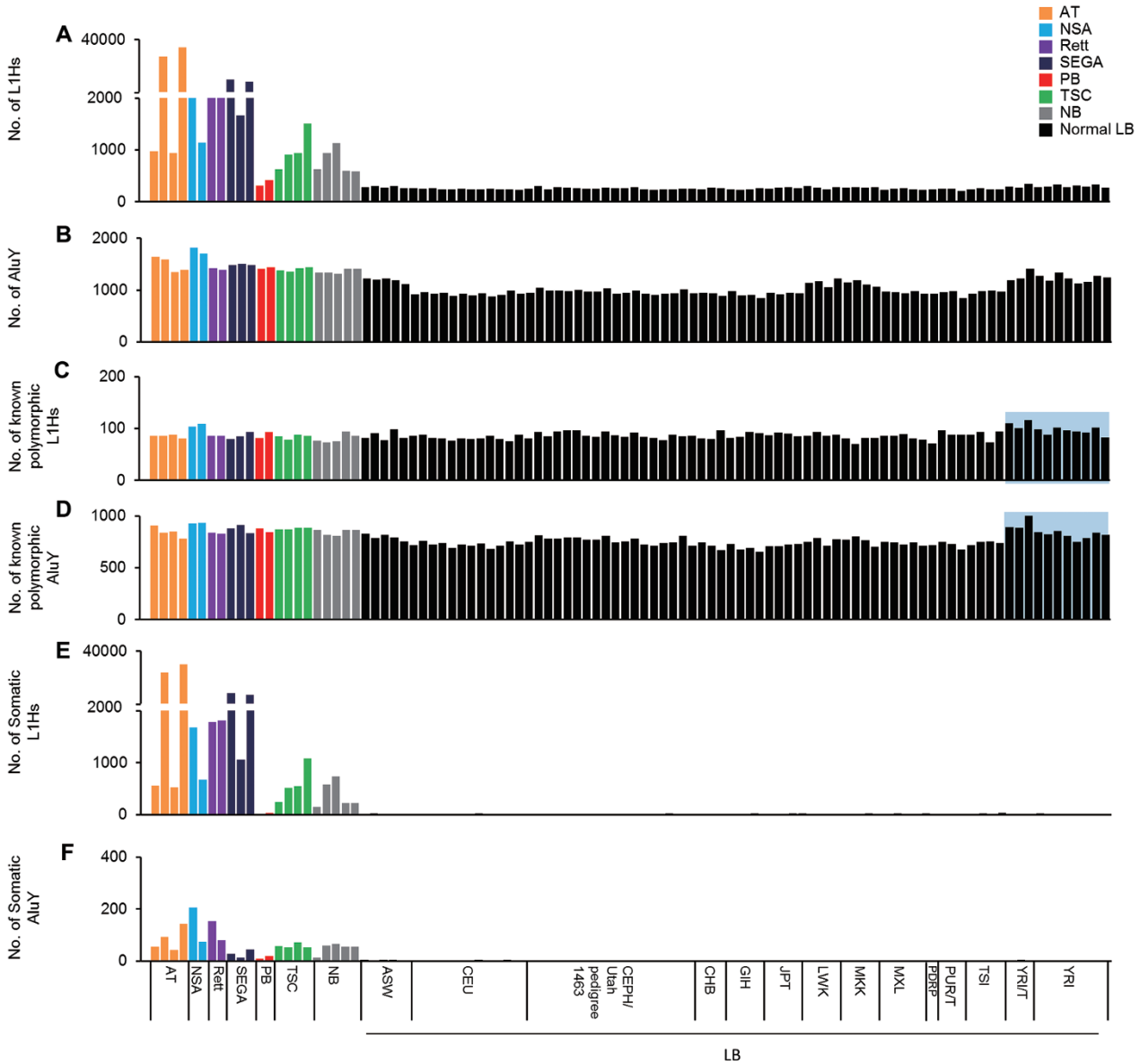
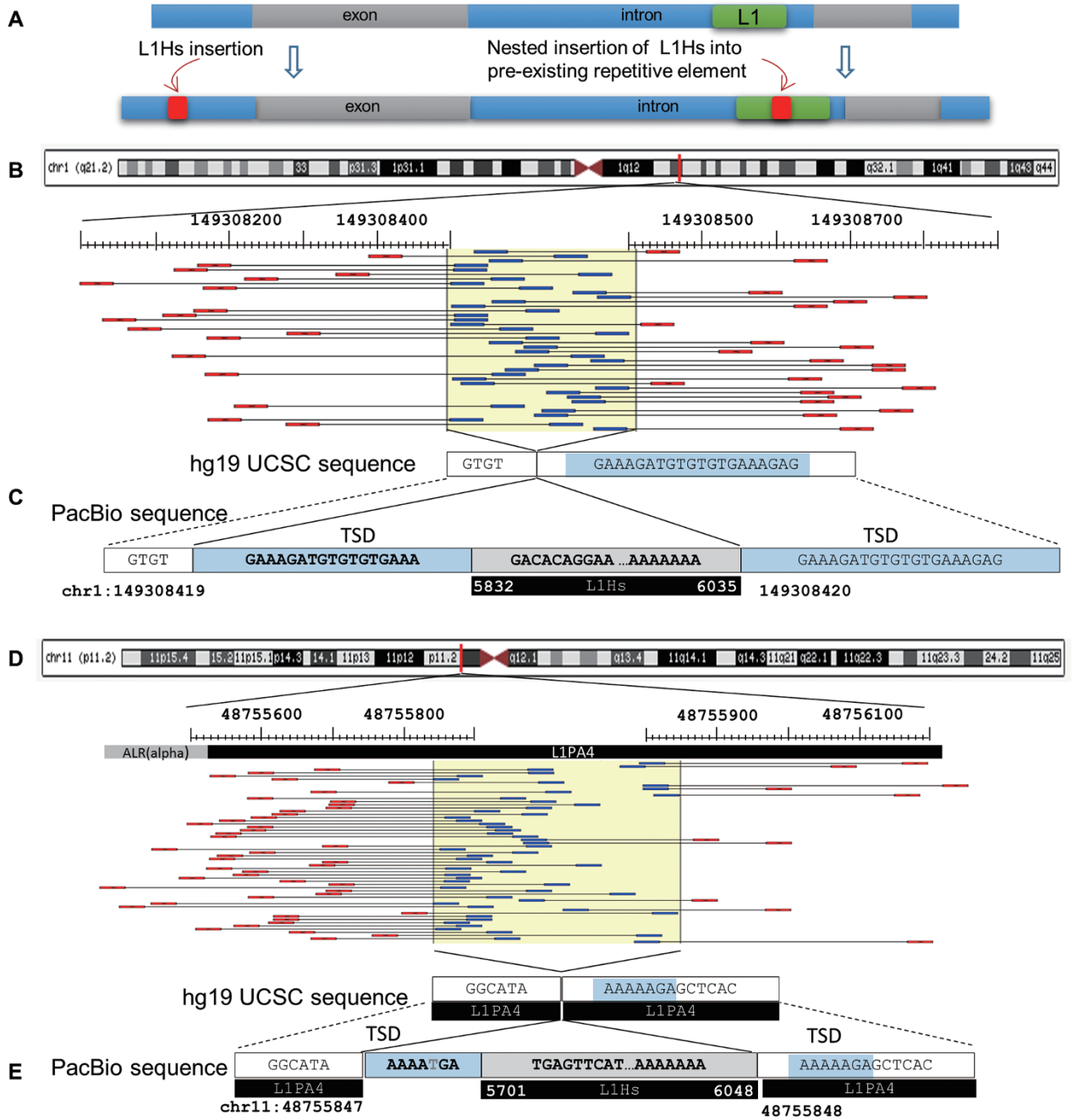


Figure 1 Novel MEI events. Novel MEI events, as detected by WGS on the CG platform, in 100 samples (detailed in Table 1). **(A)** Total novel L1Hs. High numbers were detected in the brain samples and particularly in the SEGA, Rett, AT and NSA samples. The *y* axis shows the number of novel L1Hs insertion events. **(B)** Total novel AluY. **(C**, **D)** The polymorphic novel L1Hs **(C)** and AluY **(D)** insertions we identified by CG sequencing, that were also found in the 1000 Genomes Project data. Note the uniformity of numbers of novel polymorphic germline insertions in the various samples. The normal LB samples with the highest number of polymorphic L1Hs and AluY belong to the YRI population and are highlighted in blue. **(E**, **F)** Somatic insertions of L1Hs **(E)** and AluY **(F)** are shown for all the samples we analyzed. Somatic L1Hs and AluY were high mainly in the brain samples particularly in AT, NSA, Rett and SEGA samples. Note the scale break in the *y* axis in **(A)** and **(E)** and the different scales in the different panels. The *x* axis is common to all the panels: AT, ataxia telangiectasia; NSA, non syndromic autism; Rett, Rett syndrome; SEGA, TSC-specific subependymal giant cell astrocytomas; PB, Peripheral Blood; TSC, TSC non tumoral brain sample; NB, normal brain; LB, normal lymphoblastoid cell lines representing various ethnic populations. Population name abbreviations: ASW, African ancestry in Southwest USA; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Beijing, China; GIH, Gujarati Indian in Houston, Texas, USA; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MKK, Maasai in Kinyawa, Kenya; MXL, Mexican ancestry in Los Angeles, California; TSI, Toscans in Italy; YRI, Yoruba in Ibadan, Nigeria; PUR, Puerto Rican in Puerto Rico; P (PDRP), an individual from the polymorphism discovery resource; T: (TRIO), two parents and one child.


A
exon
intron
L1
L1Hs insertion
Nested insertion of L1Hs into
pre-existing repetitive element
exon
intron
B
149308200
149308400
149308500
149308700
hg19 UCSC sequence
GTGT
GAAAGATGTGTGTGAAAGAG
C
PacBio sequence
TSD
TSD
GTGT
GAAAGATGTGTGTGAAA
GACACAGGAA ...AAAAAAA
GAAAGATGTGTGTGAAAGAG
5832
L1Hs
6035
chr1:149308419
149308420
D
48755600
48755800
48755900
48756100
ALR(alpha)
L1PA4
hg19 UCSC sequence
GGCATA
AAAAAGAGCTCAC
L1PA4
L1PA4
E
PacBio sequence
TSD
TSD
GGCATA
AAAATGA
TGAGTTCAT...AAAAAAA
AAAAAGAGCTCAC
L1PA4
5701
L1Hs
6048
L1PA4
chr11:48755847
48755848
F
Not nested
No. of somatic L1Hs
40000
20000
2000
1000
0
AT1
AT2
AT3
AT4
NSA1
NSA2
Rett 1
Rett 2
SEGA1
PB1
SEGA2
SEGA3
PB3
TSC1
TSC2
TSC3
TSC4
NB1
NB2
NB3
NB4
NB5

L1Hs retrotranspositions constitute the majority of somatic brain insertion events

We analyzed the frequency of the major retroelement types among the novel MEIs and found, as expected, that the most prevalent are the L1 and Alu retroelement families, in particular the younger subfamilies L1Hs and AluY, with some marginal representation of other retroelement families (including LTR, SVA, MER, polyA). We focused on the insertional events of the most active L1Hs and AluY elements (Figure 1).

We adopted two approaches to characterize somatic L1Hs and AluY insertion events in the brain. The first is based on the frequency of reads supporting the retrotransposition event, having in mind that insertion events represented by a read frequency of ≤ 0.3 are mostly somatic [29] (Supplementary information, Figure S5A-S5C and Data S2). To verify the authenticity of this analysis we used an orthogonal second approach that is based on the examination of pairs of matching samples from eight individuals (Table 1). Events that were detected in one of the samples from the pair and not in the other and also not reported in either the 1000G [28] or in the CG baseline MEI database (CG Public Genome Cancer Sequencing Service Guide; URLs section) are called somatic (Supplementary information, Figure S5D and S5E). A very high concordance was found between the two approaches (Supplementary information, Figure S5F); 95.7% of the events called somatic based on a read frequency ≤ 0.3 and not reported either in the 1000G [28] or in the CG baseline MEI database were also detected in one of the paired samples and not in the other. We therefore used the criterion of read frequency ≤ 0.3 to analyse somatic events in all the samples.

Figure 1E and 1F depicts the distribution of somatic L1Hs and AluY insertions in all the analyzed samples. It can be clearly seen that somatic L1Hs and AluY insertions occur almost exclusively in the brain samples and that the absolute numbers of L1Hs insertions is significantly higher than AluY insertions ($P < 0.0001$, Mann-Whitney test for both L1Hs and AluY). The number of L1Hs somatic events is 14.7 fold higher in the pathological compared to NB samples (average of 5 633 vs 382.2; $P < 0.05$, Mann-Whitney test) indicating that excessive brain somatic L1Hs retrotransposition is a common characteristic of neurodevelopmental disorders.

The majority of somatic L1Hs insertions are nested in pre-existing elements

The identification of numerous insertion events led us to systematically explore the genomic targets of retrotransposition. The distribution of the somatic insertions we identified is generally concordant with chromosome size, with a notable overrepresentation of L1Hs (12.5%) in the X chromosome (150% more L1Hs insertions than expected from chromosome size; Supplementary information, Figure S6). Interestingly, the relative germline content of L1 elements in the X chromosome is known to be the highest among human chromosomes, with one third of this chromosome consisting of L1 sequences [30]. The enrichment of somatic L1Hs insertions in the X chromosome suggests that host L1 sequences may be preferred integration sites for active L1Hs elements. To

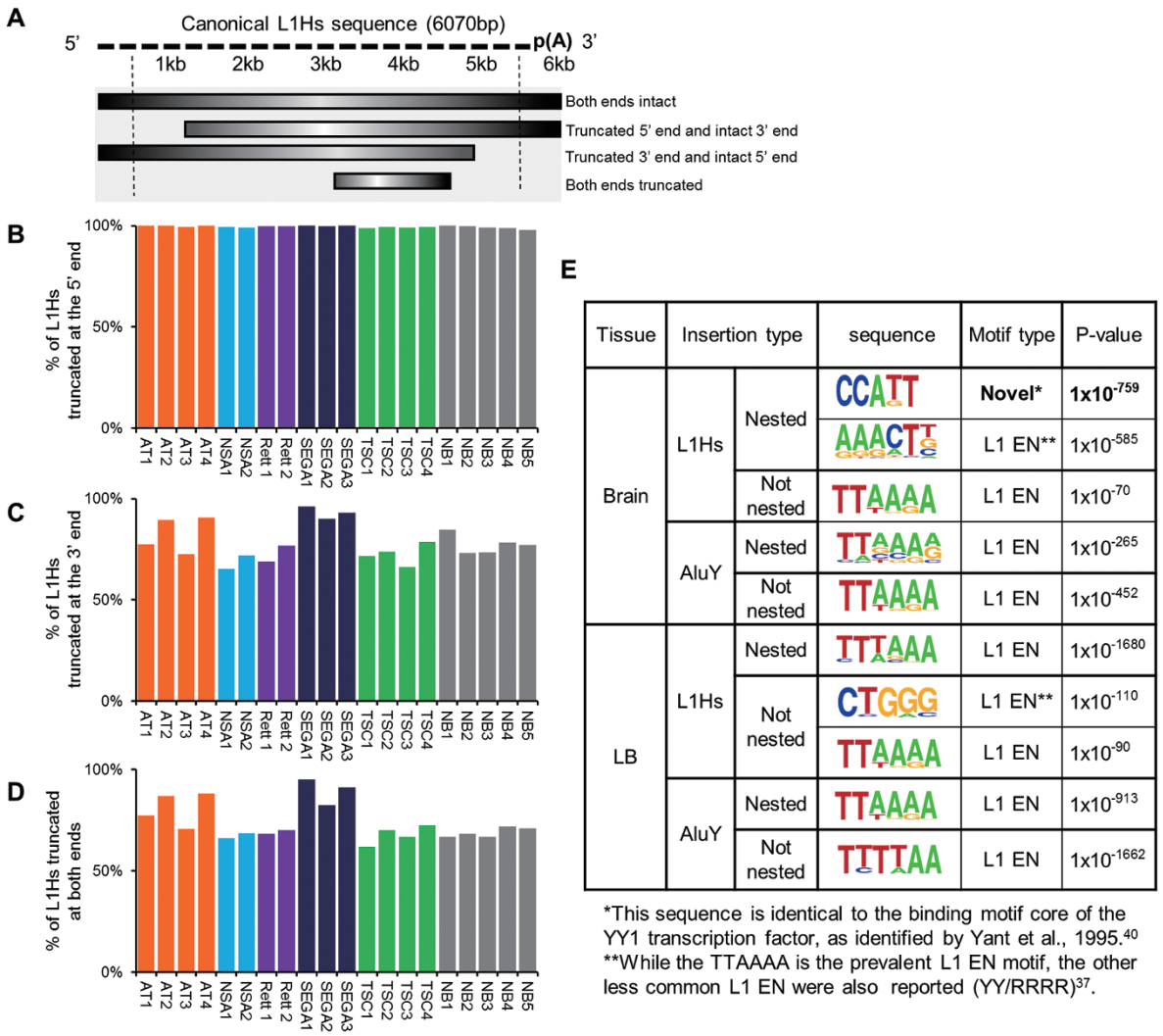
Figure 2 Structure of L1Hs insertion events as detected by CG analysis and verified by PacBio single-molecule sequencing. **(A)** Illustration of nested (right) and non-nested (left) L1Hs insertions. Nested insertions are those integrating into a pre-existing genomic repetitive element. The retrotransposon (red) can be inserted into a unique genomic region or land in a preexisting repetitive element (green). **(B**, **C)** depict a non-nested insertion event; **(D**, **E)** depict a nested insertion event. **(B)** an example of a non-nested insertion in chromosome 1. Chromosomal location of a novel L1Hs insertion identified in the SEGA1 sample is shown as a vertical red line and the genomic coordinates are shown according to the UCSC chromosome chart. Detection of the non-nested MEI was by CG paired-end reads (DNBs): red arms map to the reference genome (hg19) and blue arms map to the L1Hs canonical sequence. The arms are separated by ~400 bp spacers (black line). The region highlighted in yellow shows the L1Hs insert that is not present in the hg19 reference genome (See details in Supplementary information, Data S2). Forty out of the 270 paired end reads (DNBs) supporting this insertion event are shown here. The full coverage is given in Supplementary information, Figure S1E. **(C)** The detailed mapping by the PacBio platform of the same 203 bp long L1Hs insertion that is depicted in **(B)**. The inserted retrotransposon includes the 5 832 to 6 035 L1Hs fragment, a poly A tail and is flanked by target-site duplication (TSD) sequences (light blue boxes). **(D)** Chromosomal location and genomic coordinates of a novel L1Hs insertion nested into a preexisting LIPA4 element in chromosome 11. Fifty-seven out of the 287 supporting reads are shown. The full coverage is depicted in Supplementary information, Figure S1F. The sequence of this L1PA4 repeat in the reference genome diverged enough to represent a unique sequence, and reads could map to the insertion in this repetitive element location with high score (Supplementary information, Data S1). **(E)** The detailed mapping by PacBio of the 347 bp long L1Hs insertion nested in the L1PA4 repeat element depicted in **(D)**, includes the 5 701 to 6 048 canonical L1Hs fragment, a poly A tail and is flanked by TSDs (light blue boxes). In this case the left TSD contains a T nucleotide not present in the reference sequence (marked in grey). **(F)** Number of somatic L1Hs insertions in the analyzed samples. The *y* axis shows the number of L1Hs. The non-nested insertions in each sample are marked in yellow.

further explore this possibility we intersected the coordinates of the somatic insertions, as determined by CG WGS, with those of known genomic repetitive elements and found that the majority (93%) of somatic L1Hs insertions are nested into pre-existing repetitive elements in the genome, as illustrated in Figure 2A-2E. The bulk of the somatic L1Hs in the samples we analyzed are nested insertions (mean 83%, 54%-99%; Figure 2F). Nested L1 insertions were reported 30 years ago [31]; more than 100 L1 nested insertions were described in a study from the pre-NGS era [32] and recent studies further demonstrated that transposable elements, including L1 retroelements, are frequently embedded in pre-existing repeat elements in the germline [33, 34]. We validated the frequent occurrence of nested insertions by 3′ L1Hs capturing followed by Illumina sequencing (Supplementary information, Figure S7 and Data S3, S4); 57% of the 520 insertion events that we identified both by CG WGS and by 3′ capturing are nested. Nested insertions were also identified by Sanger sequencing as detailed in Supplementary information, Table S3A. In addition, to further substantiate the occurrence of nested insertion events by a different sequencing technology, we analyzed the very long single-molecule reads obtained by the PacBio platform that enables direct, amplification-free sequencing and mapped the nested insertions including the same polymorphic events that we identified by CG WGS (Figure 2D, 2E and Supplementary information, Figure S4E-S4H, Table S4). This analysis further supports our finding that nested insertions are prevalent and contribute significantly to genome diversity.

L1 elements are known to integrate preferentially in A/T-rich sequences and to be underrepresented in G/C rich regions [35]. We propose that pre-existing genic L1 sequences can provide an A/T-rich "landing strip" in the G/C-rich genic environment. A very high percentage (99.1%) of the somatic nested L1Hs insertions we identified integrate in the same orientation as the host repetitive sequence, as previously published for germline nested insertions [33]. The majority of nested somatic L1Hs insertions, 92.8% and 77.7% in brain and in normal LB samples, respectively, were inserted into the primate-specific L1P subfamilies. Elements of the L1PA subfamily, that have the highest homology to L1Hs (Supplementary information, Figure S8A and S8B), were found to host about 71% and 69% of the nested somatic insertions in brain and normal LB samples, respectively, while comprising only 13% of the total genomic L1 content (Supplementary information, Figure S8C). Elements of the prevalent L1MA subfamilies that are less homologous to L1Hs constitute the next most preferred integration host sequences (4.3%; Supplementary information, Figure S8). These findings indicate that the degree of homology of the host sequence to the inserted element affects integration site choice more than the abundance of the particular host element.

Features of somatic L1Hs nested insertions in the brain suggest a role for EN independent retrotransposition

Two major types of retrotransposition are known: the prevalent, classical TPRT-mediated retrotransposition that involves the L1 ORF2 encoded endonuclease (EN), and the non-classical, EN-independent (ENi) retrotransposition. The latter is characterized by the lack of an EN consensus recognition sequence at the insertion site, by frequent truncation of the inserted elements at both the 5′ and 3′ ends, and by frequent lack of target-site duplication (TSD) [36]. Transpositions of this type are considered rare events and were never reported in brain tissues. We analyzed the types of retrotransposition observed in the 84 529 L1Hs somatic insertions in our database, using stringent criteria of ≥ 10 supporting reads for each junction and of truncation occurring at least 500 bp from each end of the L1Hs canonical sequence (Figure 3A-3D). In addition to the expected frequent 5′ truncations (99.3%; 96.7%-100% of L1Hs insertions in the various samples), many of the somatic L1Hs insertions were also trimmed at the 3′ end (76.8%; 65.5%-95.6% per sample). Truncation at both the 5′ and 3′ end occurs significantly more in neurodevelopmental pathological brain samples compared to NB samples ($P < 0.022$, Student's *t*-test, heteroscedastic), in particular in the AT and SEGA samples (Figure 3D). A similar distribution of truncated L1Hs insertions was also observed when we used a more stringent analysis based on insertions supported by >30 reads at each end (Supplementary information, Figure S9A and S9B). Notably 94% of the insertions with 3′ end truncations are nested in other repetitive elements. Whereas ENi insertions with truncations at both ends were considered rare events in previous studies [12], we found that such retrotranspositions are common in brain tissues. To substantiate our findings and to verify that such insertion events are not the result of CG NGS artifacts, we looked for similar nested L1Hs insertions with both 5′ and 3′ truncations in the WGS data obtained by the PacBio technology. We provide 21 examples of such L1Hs insertions, out of 23 events of L1Hs nested insertions in pre-existing L1 family elements (Supplementary information, Table S4). Our data shows that in the majority of truncated insertions, the 3′ end is located between positions 4 500 and 5 500 of the canonical L1Hs sequence, similar to a previous report [37]. Such 3′ truncated inserted elements are therefore predicted to escape detection by the 3′ end-capturing methods used in most



E

Tissue	Insertion type		sequence	Motif type	P-value
Brain	L1Hs	Nested	CCATT	**Novel***	**$1x10^{-759}$**
			AAACTG	L1 EN**	$1x10^{-585}$
		Not nested	TTAAAA	L1 EN	$1x10^{-70}$
	AluY	Nested	TTAAAA	L1 EN	$1x10^{-265}$
		Not nested	TTAAAA	L1 EN	$1x10^{-452}$
LB	L1Hs	Nested	TTTAAA	L1 EN	$1x10^{-1680}$
		Not nested	CTGGG	L1 EN**	$1x10^{-110}$
			TTAAAA	L1 EN	$1x10^{-90}$
	AluY	Nested	TTAAAA	L1 EN	$1x10^{-913}$
		Not nested	TTTTAA	L1 EN	$1x10^{-1662}$

*This sequence is identical to the binding motif core of the YY1 transcription factor, as identified by Yant et al., 1995.[40]
**While the TTAAAA is the prevalent L1 EN motif, the other less common L1 EN were also reported (YY/RRRR)[37].

Figure 3 Structural characterization of inserted L1Hs insertions and enriched motifs at L1Hs and AluY insertion sites. **(A)** Schematic illustration of the main types of novel L1Hs insertions: The coordinates of the canonical 6 070 bp long L1Hs sequence (broken line) are presented. The bars depict examples of intact and truncated inserted elements. Insertion types were defined based on stringent criteria: ≥ 10 supporting reads (DNBs) for each end, shortening of > 500 bp from the canonical 5′ end and of > 570 bp from the canonical L1Hs 3′ end (broken vertical lines). **(B)** The percentage of L1Hs insertions truncated at the 5′ end, in the brain samples. **(C)** The percentage of L1Hs insertions truncated at the 3′ end, in the brain samples. **(D)** The percentage of L1Hs insertions truncated at both 5′ and 3′ ends, in the brain samples. **(E)** The most enriched motifs detected at the insertion sites of somatic and germline L1Hs in the brain and in the normal LB samples. In all groups, except the brain nested L1Hs insertions (where a novel motif was identified), the most enriched motif is the known L1 EN cleavage sequence (YY/RRRR) [38].

previous studies [3, 13, 14].

It was suggested that the loss of the ATM protein as occurs in AT patients may lead to longer L1 insertions [20]. Indeed, in our study, the length of the insertions in the AT samples was moderately longer (Supplementary information, Data S5 and Figure S10).

To further characterize brain retrotranspositions, we studied the L1Hs and AluY integration site ranges, which are defined as the reference coordinates specifying the start and end points of the region where the insertion event is likely to reside (Cancer Sequence Service Data File Formats, see URLs section). We found that L1Hs and AluY insertion regions are A/T-rich (> 64% and > 58%, respectively) as reported before [35]. Classic EN-dependent insertions occur predominantly in the TTTT/AA consensus sequence and its variants [38],

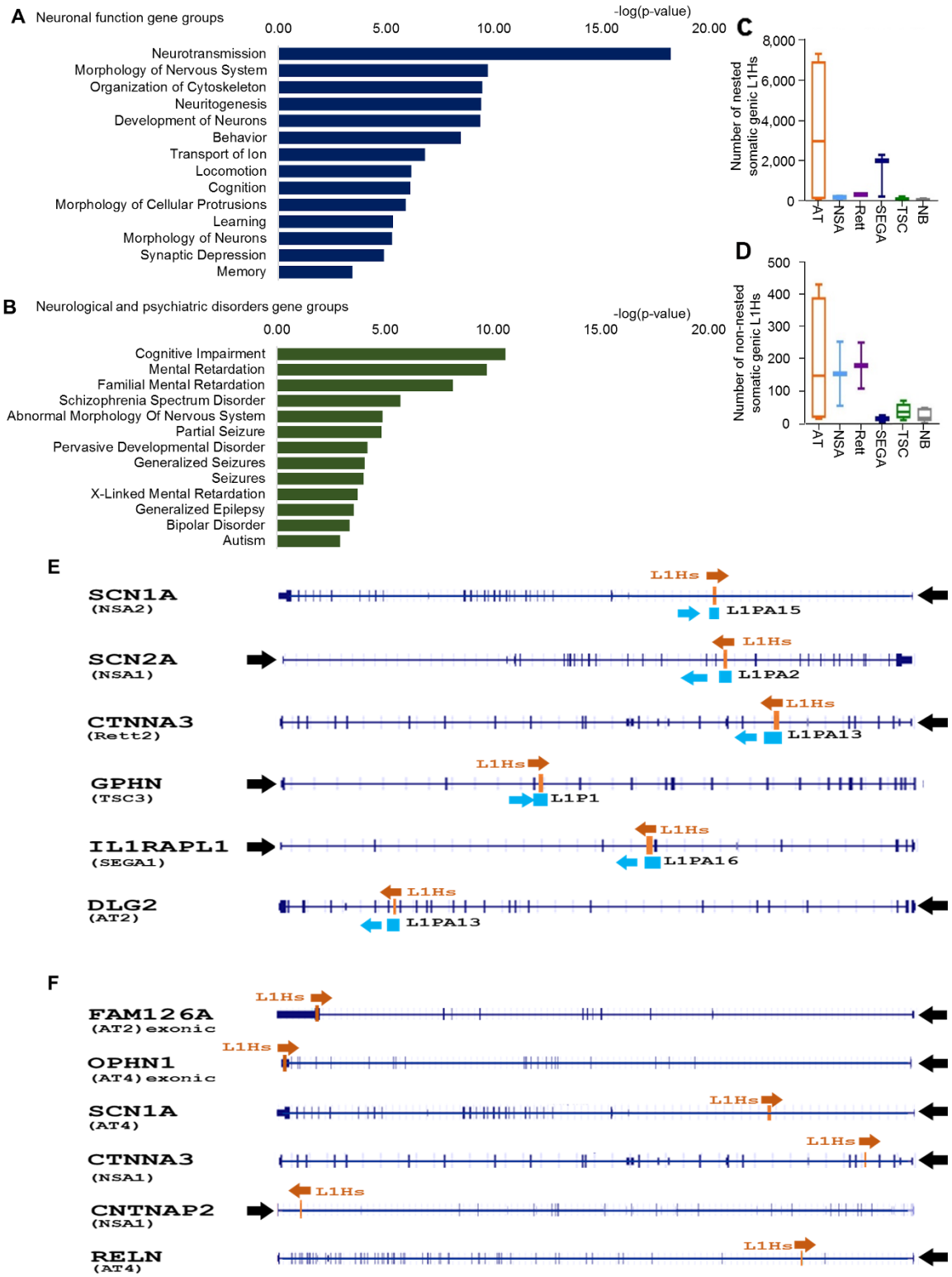
whereas in ENi retrotranspositions this consensus sequence is not found [36, 37, 39]. We applied an unbiased motif-search tool (HOMER, URLs section) on L1Hs and AluY and compiled 50 bp insertion ranges of nested and non-nested integrations in the brain samples and in the normal LB samples. We found that the most enriched consensus sequences in the various categories were consistent with the TTTT/AA L1 EN motif ($P = 1 \times 10^{-70}$ – 1×10^{-1680}; Figure 3E); Strikingly, a novel CCATT motif, not known as a L1 EN consensus sequence, is highly enriched in nested brain L1Hs insertions ($P = 1 \times 10^{-759}$; Figure 3E). This sequence is similar to the binding motif core of the YY1 transcription factor [40], that was shown to be recruited to DNA damage sites [41] and was also suggested to be involved in the regulation of L1Hs transcription [42]. As this motif is enriched specifically in nested insertions we looked for this consensus sequence in all genomic L1 sequences that are potential targets for retrotransposition and found it to be enriched 2 fold ($P < 10^{-16}$, χ^2 test) compared to non-L1 sequences. Furthermore, it is four times more prevalent in the L1PA subfamily, the preferred target of nested L1Hs in all the brain samples, compared to other L1 family elements ($P < 10^{-16}$, χ^2 test). These findings suggest that a non TPRT-EN-mediated mechanism is the main player in brain retrotransposition. This ENi mechanism is known to be associated with DNA damage and repair and has not been previously described in neural tissues.

Genic distribution and functional analysis of L1Hs insertions

Most somatic genic insertions in the brain samples involve L1Hs elements (21 603), while only 621 AluY genic insertions were found. We therefore focused on the distribution and function of the L1Hs elements. Genic insertions comprise 25% of the total somatic events and most of them were shown to be nested (89%). The majority of L1Hs genic insertions are located in introns with only 117 affected exons, half of them not nested. Careful inspection of the exonic insertions revealed that the retroelements are mainly located in the 5′ and 3′ UTR exons. Genic L1Hs are preferentially inserted in an antisense orientation to the host gene (67%; 40%-84%). A similar bias towards antisense genic L1 insertions was reported before, where it was shown that this orientation inhibits gene transcription by producing premature polyadenylation [43]. Analysis of insertions into microRNA sequences (as listed in the hg19 UCSC), did not reveal L1Hs somatic insertions in these sequences. On the other hand 2 524 long non-coding RNAs (LncRNA; as listed in the hg19 UCSC) had somatic insertions of L1Hs (Supplementary information, Figure S11A); most of these insertions were nested (93.4%; Supplementary information, Figure S11B).

There is a significant correlation between the number of pre-existing L1 elements per gene to the gene length ($r = 0.76$; $P < 2 \times 10^{-100}$). The number of somatic L1Hs insertions per gene is correlated with gene length ($r = 0.6$; $P < 2 \times 10^{-100}$) and with the number of pre-existing L1 elements per gene ($r = 0.57$; $P < 2 \times 10^{-100}$). This is in agreement with the finding that the vast majority of L1Hs somatic insertions in genes are nested in pre-existing repetitive elements.

Functional group analysis of genes affected by somatic insertions of L1Hs revealed that several groups of neuronal function are significantly enriched, mainly in the SEGA, Rett, AT and NSA samples. Figure 4A and 4B present the significantly enriched functional categories (Ingenuity IPA; ≥ 20 genes and $P \leq 5.0 \times 10^{-3}$; Fisher's exact test; Benjamini-Hochberg). Neuronal function groups enriched for genes with L1Hs insertions include "neurotransmission", "morphology of nervous system", "neuritogenesis", "behavior", "cognition" and "learning" (Figure 4A). In Figure 4B, we present the significantly enriched categories of neurological and psychiatric disorders such as "cognitive impairment", "mental retardation", "schizophrenia spectrum", "seizures" and "autism". The finding of these enriched functional groups is in accord with the hypothesis that L1 retrotransposition contributes to the diversity and the plasticity of neuronal circuits and with the emerging understanding that uncontrolled L1 insertion is involved in various neuropathologies [15, 16]. Figure 4C and 4D depict the distribution of the total and the non-nested genic insertions, respectively, in the various samples representing various neurodevelopmental brain disorders and the NB. It can be seen that the highest numbers of non-nested somatic L1Hs insertions are found in AT, NSA and Rett samples while the TSC related samples (SEGA and non-tumorous brain tissue) as well as the NB have fewer non-nested genic somatic insertions. There is no direct correlation between the total number of retrotransposition events and the number of non-nested insertions. About 3 000 genes are affected by somatic L1Hs insertions; only one third of these are affected by non-nested insertions. Yet, the functional categories in the two gene groups are similarly enriched (Figure 4A, 4B and Supplementary information, Figure S12). A representative list of neural genes with high somatic L1Hs insertion numbers in brain samples is given in Supplementary information, Table S5. Selected examples are depicted in Figure 4E, 4F. The list includes known neural genes that are associated with neurological and psychiatric disorders such as autism, epilepsy and schizophrenia. For example, we found 117 somatic L1Hs


A
Neuronal function gene groups
-log(p-value)
0.00
5.00
10.00
15.00
20.00
Neurotransmission
Morphology of Nervous System
Organization of Cytoskeleton
Neuritogenesis
Development of Neurons
Behavior
Transport of Ion
Locomotion
Cognition
Morphology of Cellular Protrusions
Learning
Morphology of Neurons
Synaptic Depression
Memory
B
Neurological and psychiatric disorders gene groups
-log(p-value)
0.00
5.00
10.00
15.00
20.00
Cognitive Impairment
Mental Retardation
Familial Mental Retardation
Schizophrenia Spectrum Disorder
Abnormal Morphology Of Nervous System
Partial Seizure
Pervasive Developmental Disorder
Generalized Seizures
Seizures
X-Linked Mental Retardation
Generalized Epilepsy
Bipolar Disorder
Autism
C
Number of nested somatic genic L1Hs
8,000
6,000
4,000
2,000
0
AT
NSA
Rett
SEGA
TSC
NB
D
Number of non-nested somatic genic L1Hs
500
400
300
200
100
0
AT
NSA
Rett
SEGA
TSC
NB
E
SCN1A
(NSA2)
L1Hs
L1PA15
SCN2A
(NSA1)
L1Hs
L1PA2
CTNNA3
(Rett2)
L1Hs
L1PA13
GPHN
(TSC3)
L1Hs
L1P1
IL1RAPL1
(SEGA1)
L1Hs
L1PA16
DLG2
(AT2)
L1Hs
L1PA13
F
FAM126A
(AT2)exonic
L1Hs
OPHN1
(AT4)exonic
L1Hs
SCN1A
(AT4)
L1Hs
CTNNA3
(NSA1)
L1Hs
CNTNAP2
(NSA1)
L1Hs
RELN
(AT4)
L1Hs

insertions in the CTNNA3 gene, in which exonic deletions have been previously shown to segregate with NSA [44]. A total of 62 insertions were found in the *GPHN* gene, where exonic deletions were suggested to increase the risk of autism, schizophrenia and seizures [45]. In total, 37 somatic L1Hs insertions were identified in the *IL1RAPL1* gene in our study. Mutations in this gene were previously associated with autism and cognitive impairment [46]. We also found insertions in several of the pathological brain samples in the *SCN1A* and *SCN2A* genes that were recently reported as possible genetic drivers of autism [47, 48]. Finally, the *DLG2* gene which was repeatedly shown to be mutated in schizophrenia, and was recently found to be a target for somatic L1-associated variants [13], had 141 somatic insertions in the brain samples analyzed (Supplementary information, Table S5). Figure 4E and 4F provide examples of somatic genic insertions in selected neuropathology related genes. These comprise nested and non-nested insertions, including exonic non-nested events, which are predicted to have deleterious effects.

Discussion

Somatic mutations are usually more detrimental than inherited mutations, as they escape evolutionary selection [49]. Yet, unique somatic mutation mechanisms can provide functional and survival advantages as exemplified by the somatic hyper mutation machinery that diversifies the immunoglobulin genes [50]. By analogy, retrotransposition-facilitated somatic mutagenesis is suggested to contribute to neuronal diversity and plasticity [14-16, 27]. We compiled a database of unprecedented magnitude containing more than 80 000 somatic L1Hs insertions in the brain and identified unique characteristics of brain retrotransposition that were not reported in previous studies of brain mobile elements.

L1 somatic insertions can alter the expression of brain genes by a variety of mechanisms [1, 27, 51, 52]. Our findings allude to defense mechanisms that favor L1-mediated fine-tuning of neural circuits and plasticity over detrimental insertions [14-16, 27]. We demonstrate that only a quarter of the somatic L1 insertions occur in genic regions and that the involvement of exons, and coding sequences in particular, is exceptionally rare, hence lessening the risk of damaging insertions. Furthermore, the nesting of most somatic L1Hs within pre-existing repetitive elements constitutes a "lightning rod" that shields genes from harmful insertional mutagenesis. The presence of conserved genomic regions devoid of repetitive elements, such as the hox-genes loci, indicates that germline occurrence of repetitive elements is evolutionarily selected to prevent damaging outcomes [2, 53]. The nesting of somatic L1Hs elements in tolerant, evolutionarily selected, repetitive sequences is similarly predicted to protect against harmful effects [33].

The brain samples analyzed here were obtained from individuals with neurodevelopmental disorders that are characterized by extensive inter- and intra-family variability of clinical features. Clinical heterogeneity was described in family members, and even among identical twins affected by TSC, Rett and NSA, carrying the same disease-causing mutated gene [54-56]. Increased stochastic retrotransposition in the brain emerges as a common pathogenic mechanism that is central to the clinical diversity observed in these disorders. Recent studies, based on the analysis of retrotransposon expression or on the quantitative PCR of genomic DNA, suggest that mobile element insertions are also relevant to neurodegenerative disorders such as amyotrophic lateral sclerosis and frontotemporal lobar degeneration, psychiatric diseases such as schizophrenia and aging [57, 58]. Aberrant retrotransposition therefore emerges as a common feature of diverse brain neuropathologies. We report here for the first time that retrotransposon insertions are prevalent in the unique TSC-archetypal SEGA brain tumors [24]. SEGAs arise in the subventricular zone where postnatal neurogenesis occurs, and they exhibit a mixed glioneu-

Figure 4 Genes with somatic L1Hs insertions. **(A**, **B)** Enriched functional groups of genes with L1Hs insertions. Genes that were inserted with at least one somatic L1Hs event were analyzed for functional group enrichment. Enriched neuronal function gene groups **(A)** and neurological and psychiatric disorder gene groups **(B)** with more than 20 genes and $P \leq 5.0 \times 10^{-3}$ (Ingenuity IPA; Fisher's exact test; Benjamini-Hochberg) in the brain samples analyzed. **(C**, **D)** Box plot charts of the number of nested **(C)** and non-nested **(D)** somatic L1Hs insertions in genes as distributed in each of the various neurodevelopmental disorders and the normal brains: AT (n = 4), NSA (n = 2), Rett (n = 2), SEGA (n = 3), TSC (n = 4) and NB (n = 5). Non nested L1Hs insertions are more abundant in the AT, NSA and Rett pathological samples. **(E**, **F)** Examples of genes associated with neurological and psychiatric disorders with nested **(E)** and non nested **(F)** L1Hs somatic insertions. The L1Hs insertions examples (orange bars) are located in introns (exons are marked as dark blue bars) and are nested **(E)** in pre-existing L1 family members (blue box) in the same orientation as the host element (orange and blue arrows for the novel L1Hs and the host L1, respectively), and opposite to the gene orientation (black arrows) in most of the cases. The non-nested L1Hs insertions **(F)**, are also inserted opposite to the gene orientation. The non-nested insertions in FAM126A and OPHN1 are examples of exonic insertions.

ronal phenotype [24]. These characteristics and the study of genetically manipulated TSC mice [25], indicate that SEGAs originate in NPCs. Based on the demonstration of physiologic L1 retrotransposition in genomes of NPCs [15, 27], we speculated that transformation of such cells, "frozen" in the time/differentiation window of permissive L1 retrotransposition, could result in a tumor whose genome is invaded by multiple insertions. Indeed, the number of somatic insertions in the NPC-derived SEGA tumors we analyzed is larger than that reported in any tumor studied so far [11].

We found that neuronal genes and genes associated with a variety of neurologic and psychiatric disorders are preferred targets of somatic brain retrotranspositions, as was shown before [14]. This enrichment of insertion events was suggested to reflect open chromatin in genomic regions expressing the brain-specific genes and also the fact that neural genes tend to be longer than average [27]. We show here that the genic density of pre-existing L1 elements is an additional factor favoring insertion in neuronal genes.

We found that the majority of nested L1Hs insertions in the brain have characteristics of non-classical retrotransposition, occur preferentially in A/T rich L1 sequences, and are found in the same orientation as the host repetitive element. The ENi retrotransposition was revealed initially from the analysis of L1 retrotransposition in cells lacking components of the non-homologous end joining pathway of DNA double-strand break (DSB) repair [36] and later in cells with intact repair mechanisms [37]. Previous studies of ENi L1Hs insertions [36, 37] detailed them as having 3′ truncations, diverging significantly from the L1-EN consensus, and lacking TSDs and a poly(A) tail. These criteria characterize the majority of somatic insertions in the samples we analyzed. The last two features are not reported in this study, as the CG method enables the detection of insertion events location in a range of several nucleotides. This methodology is therefore inadequate for the precise detection of TSD sequence at the junctions of the insertion events. We report, however, several TSD sequences of validated events using Sanger or PacBio methodologies. In some verified insertions, we found that indeed, TSD sequences were missing. The findings of insertion events with no TSD further allude to the ENi mechanism of insertion. A similar integration mode was shown to be involved in RNA-mediated DNA repair [36, 59]. Recently, DSBs introduced into mouse zygotes by the CRISPR/Cas system were shown to be repaired by retrotransposon sequences [60], further supporting the occurrence of DNA breakage-related retrotransposition. Such alternative retrotransposition events were also demonstrated by recent studies of tumours by WGS [12, 61]. Frequent DNA breaks were demonstrated in normal mouse brains, mainly in the dentate gyrus where enhanced retrotransposition normally takes place, and were suggested to be associated with brain function [62]. DNA breaks were also linked to brain pathology, as mice transgenic for human amyloid precursor protein were shown to have increased neuronal DSBs [62]. An increased load of DNA breaks was reported also in aging and in Alzheimer brains [63] and the XRCC4 and ligase 4 components of the DSB repair machinery were shown to be essential for NPC survival during neurogenesis [64]. It was shown recently that recurrent DNA breaks accumulate in long genes that are involved in synapse function, neural cell adhesion and mental disorders [65], categories similar to the gene categories found in our study to be enriched for L1Hs insertions. Interestingly, we observed the highest insertion burden in brain samples of patients suffering from AT, the prototypic DNA repair disease, further supporting a link between DNA damage and retrotransposition [20, 22]. This goes in line with the novel CCATT motif we identified at insertion sites of somatic L1Hs retrotransposition in the brain that may be associated with the YY1-binding sequence that has relevance to both DNA damage and L1 promoter activation [40, 41].

Our work expands the spectrum of neuronal disorders with WGS-proven enhanced retrotransposon insertions to include Rett, AT, TSC as well as TSC-related SEGAs and NSA. We speculate that in these disorders, increased DNA damage and enhanced retrotransposition overcome the "lightning rod" safeguard mechanism, culminating in damaging somatic insertions.

Materials and Methods

Human samples

In this study we analyzed the following 100 samples: Three subependymal giant cell astrocytomas; (SEGA, patients1-3), two matching blood samples (PB, patients1 and 3), five NB samples (NB1-5) and brain tissues of patients affected by the following clinical syndromes: tuberous sclerosis complex (TSC1-4), Rett syndrome (Rett1-2), ataxia-telangiectasia (AT1-4) and non-syndromic autism (NSA1-2). The SEGA and PB samples were obtained with the approval of the institutional IRB. The NB, TSC, Rett, AT and NSA brain tissues were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD. The data of 78 lymphoblastoid cell lines from normal individuals originating from 11 different ethnic populations (described in Supplementary information, Table S1A and S1B) are from the Complete Genomics public genome data base (see "Complete Genomics public genome data", URLs section). In addition we used the human 54× data set from the PacBio web site (See "The Pacific Biosciences web site", URLs section).

DNA extraction

DNA was extracted from frozen brain tissues using Genomic DNA Lysis buffer (0.4 M NaCl, 1% SDS, 0.1 M Tris-HCl pH 8.1, 5 mM EDTA, 0.05% proteinase K) and purified with phenol:chloroform:isoamyl alcohol (PCI 25:24:1, Invitrogen, cat 15593-031) using PLG tubes (MaXtract, Qiagen, cat 129016). DNA was extracted from the blood samples using G-DEX™ IIc (Genomic DNA Extraction Kit, cat 17231). The yield and the quality of extracted DNA were assayed using a NanoDrop and a Qubit 2.0 Fluorimeter (Invitrogen, dsDNA HS Assay, cat Q32854). DNA integrity was checked by electrophoresis in 1% agarose gel.

CG whole-genome sequencing

WGS was performed by unchained base reads on self-assembling DNA nanoarrays (Complete Genomics, GC), detailed in the Supplementary information, Figure S1A. Briefly, genomic DNA was fragmented by sonication to a mean length of 400 bp and fragmented DNA was treated to create non phosphorylated blunt termini. The end-repaired DNA fragments are then ligated to the synthetic adaptor1-4 (Ad1-Ad4) with a novel nick translation ligation process that produces efficient adaptor fragment ligation with minimal fragment-fragment and adaptor-adaptor ligation. The purified adaptor-ligated material was subjected to selective amplification of the template containing both left and right adaptor arms. The DNA was then subjected to conditions allowing intramolecular hybridization (circularization) and amplification of the circular element to give a "DNA Nano Ball" (DNB). Each DNB (CG paired-end reads) therefore comprised an amplified circular segment representing a genomic location with a 400 bp gap between the left and the right arms. CG uses combinatorial probe anchor ligation (cPAL) for sequencing. The method uses hybridization of an anchor molecule to unique adapters. Labeled degenerate 9-mer oligonucleotides with specific fluorophores for each nucleotide were used. Sequencing was performed by hybridizing a specific probe to a template followed by ligation to an anchor using T4 DNA ligase and imaging. This process was repeated and the adjacent nucleotides determined [66, 67].

Characterization of L1Hs and AluY insertions by CG WGS of the 100 samples analyzed

The MEI data set of the 100 samples that were sequenced by CG WGS (Table 1 and Supplementary information, Table S1) was analyzed and the insertion sites characterized using Spotfire DecisionSite for Functional Genomics v. 9.1.2™ (Somerville, MA) and Partek Genomic suite V.6 to generate the database allowing filtering and statistics. Bedtool's IntersectBed tool (URLs section) was used for the characterization of the genomic location of the insertions (hg19). Ingenuity IPA (version 26127183) was used for functional analysis of inserted genes.

Analysis of whole-genome sequencing reads and identification of MEI events by Complete Genomics whole-genome sequencing

WGS was performed by Complete Genomics (CG) technology using unchained base reads on self-assembling DNA nanoarrays (Supplementary information, Figure S1A) [66, 67], which allows efficient mapping of repetitive elements in the genome. WGS reads were analyzed using the CG analysis suite (CGA), as detailed in the CG manual ("Data File Formats - Cancer Sequencing Service", and "CGA" see URLs section), which includes a chapter regarding mobile element analysis.

CG sequence determination methodology is based on unchained sequencing of DNA by combinatorial probe anchor ligation (cPAL) [66]. This approach is based on the detection of ligation products of an anchor oligonucleotide that is hybridized to an adaptor sequence. A fluorescent degenerate sequencing probe containing a specified nucleotide at the interrogation position was used for the detection of the specific position of each nucleotide. Each read position had a separate pool of probes. The repetition of the process enables the identification of successive nucleotides [67].

The schematic illustration of the sequencing methodology and its advantages for the analysis of repetitive elements is given in Supplementary information, Figure S1A-S1D. To detect MEI events, using the CG methodology, we identify mate pairs in which one end maps outside the inserted element and the other end maps within the inserted element. Thus, we look for clusters of mate pairs in which one of the ends maps uniquely to the reference genome while the other end maps to a sequence that is ubiquitous in the reference, namely, a repetitive element. The cluster of uniquely aligned arms acts as an anchor for the reference genome location of the insertion on both sides. Since the arms are separated by about 400 bp, the second arm is expected to align to the reference sequence at a distance of about 400 bp. When the predicted alignment pattern is observed, no insertion event is present at this location. In case that there is no such alignment, the sequence of the second arm can be mapped to a specific repetitive element by alignment to the hg19 RepeatMasker database, containing Alu, LINE, HERVK, LTR, MER and SVA element sequences. The cluster of the aligned reads determines the detailed structure of the inserted element including the type of element and the start and end coordinates relative to the canonical sequence of the specific repeat element. To determine a MEI event, one mate must map uniquely, and the other mate needs to map to a mobile element that is not at that site in the reference genome.

Since the insertion location by the CG methodology is given as a range of insertions, the determination the TSD is not precise. Therefore, we do not report these sequences.

For the characterization of a nested insertion we map each mate to the reference genome. If one mate maps uniquely, it anchors the mate-pair at that position. This unique location might be a pre-existing repetitive element. Importantly, many mobile elements in the genome have sufficiently diverged to the extent that at least part of their sequence is unique, allowing definite localization. To assess the ability to align insertions to repetitive sequences in the genome, we performed a simulation and re-alignment of L1-derived reads in the human genome (Supplementary information, Data S3). This simulation shows that 76% of the CG reads can be unequivocally mapped back to the single-unique location in the genome. In addition, we performed an independent analysis of selected MEI calls using an in-house script to visualize the CG reads in both ends of the insertion sites, as well as the DNB (CG paired-end reads) arms that were aligned to the L1Hs sequence. Visualizations of the CG reads supporting both non-nested and nested L1Hs insertions are exemplified in Figure 2B and 2D, respectively, and in Supplementary information, Figure S1E and S1F. Our detailed in-house "analysis and visualization support for Complete Genomics MEIs" procedure is given in Supplementary information, Data S4. More examples of novel L1Hs insertion events identified in SEGA1

and PB1 are visualized in http://www.sheba-cancer.org.il/l1/view_insertions.html. In this visualization we show clusters of aligned DNBs (CG paired-end reads) that support insertion events. Red rectangles represent the arm that is aligned to a unique hg19 reference genome sequence and anchor the insertions at their genomic location, the blue rectangles represent the arms that could not be uniquely aligned to the hg19 reference genome but could be aligned to the L1Hs canonical sequence. Clicking on the rectangles representing the arms opens a data card with the exact alignment and scores.

In our work, we report only insertions that fulfill three conditions: (1) insertion length ≥ 45 bases; (2) the insertion events supported by at least 10 reads (DNBs); (3) insertion score (Phred-like score) ≥ 50. In addition, we merged insertion events from the same element type that were closer than 35 bp apart in order to call each event once. CG reports also insertion events that have no sufficient information regarding the number of reads (DNBs) supporting the reference genome; these events were not included in our analysis (See Supplementary information, Figure S2 for the overall study workflow). The CG analysis suite (CGA; see URLs section) was used for the samples analyzed in this study. All the samples we sequenced were analyzed with CGA 2.5 excluding the SEGA, the PB and the NB1 samples that were analyzed with CGA 2.0. Fifty-one out of the 78 samples deposited in the CG public database that were used for our analysis were analyzed by both versions. We performed a concordance test with insertions reported in control LB samples from both versions (LB1-7; LB10-11; LB16; LB30-33; LB35-LB48; LB50-59; LB61-62; LB64-69; LB71-72; LB74; LB76; LB78). Concordant events are insertions reported as the same element in both analyses at the same location (insertion ranges intersected). 96% and 97% of the unfiltered and filtered insertions, respectively, were concordant. The values given for the samples analyzed with both versions are of the CGA 2.5 version.

Characterization of nested insertions by CG WGS

Nested insertion events are determined here by the intersection of the insertion range of an event (median insertion range length is 96 bp and 46 bp for L1Hs and AluY, respectively) with the location of repetitive elements in the reference genome (hg19) as reported by the RM. Events were called nested insertions when at least 90% of the insertion range intersected with a repetitive element [68]. The use of this stringent criterion is expected, therefore, to underestimate the reported number of nested insertions. We detected many insertions that are located in existing repeat elements (nested insertions). The detection of such insertion into repetitive elements relies on the mapping to unique sequences in the preexisting repeat element in the reference genome that has diverged enough during evolution. In order to support the detection of such nested MEIs, we tested whether segments originating from L1 family sequences can be uniquely mapped to the reference genome as described in Supplementary information, Data S3.

Characterization of somatic insertions by CG WGS

Somatic insertion determination is detailed in Supplementary information, Data S2.

Verification of L1Hs insertion events detected by CG WGS using L1Hs 3′ targeted sequencing

L1Hs 3′ end flanking regions were captured using hemi-specific PCR followed by Illumina NGS, as described before [3]. We amplified 3′ end flanking regions of human-specific L1Hs insertion sites of the three SEGA samples (SEGA1-3) and of a NB sample (NB1). Targeting of L1Hs insertions was performed using two sets of primers; one set designed to complement the 3′ end of the human-specific L1Hs element using both 3′ proximal primer and 3′ distal primer. The second set included degenerate random primers for the flanking genomic regions downstream of the L1Hs insertions. Amplicons were designed to include Illumina's sequencing library adaptors. The amplified regions were then sequenced with Illumina (72 bp length reads) to produce reads with an opposite orientation to the captured L1Hs. Novel insertion events were identified using a bioinformatic approach adapted from Ewing *et al*. [3]. We trimmed 10 bp from the 5′ end and 18 bp from the 3′ end in order to avoid mismatches that may be caused by partial matching of the degenerate primers at the 5′ end and the low quality reads at the 3′ end. We used Bowtie [69] for alignment to the human reference genome (hg19) and MergeBed (URLs section) for read clustering. In order to identify novel L1Hs insertions, including nested events, the clusters were filtered to omit all the L1Hs in the reference genome (RM) containing the specific sequence used for the capture procedure:

The reads include both novel and known elements reported in the reference genome. In order to identify novel L1Hs insertions, including nested events, the clusters were filtered to omit all the L1Hs in the reference genome (RM) containing the specific sequence used for the capture procedure. The bioinformatic analysis enabled the detection of L1Hs insertions and the filtering-out of elements existing in the reference genome (hg19 RM) that could be targeted by both the 3′ end primers we used. The novel L1Hs insertions were identified by searching the complete human genome (hg19), with the Bowtie alignment tool [69] using the -a option, which retrieves all hits found. The primer sequences for this search were: 5′-GGGAGATATACCTAATGCTAGATGACAC-3′ (3′ distal primer) and 5′-GTACCCTAAAACTTAGAG-3′ (3′ proximal primer) [3]. Hits detected by both primers and separated by less than 100 bp in the correct orientation (opposite to the reference) and order (when the most 3′ distal primer is downstream to the 3′ proximal primer) were filtered out, using the closestbed program of BEDTools [68] using the -d and -s options.

L1Hs insertion validation by Sanger sequencing

Sanger sequencing of L1Hs insertion junctions was performed on PCR products for each event, using primers from the inserted L1Hs sequence and from the neighboring genomic region, and carried out using BigDye Terminator Cycle Sequencing (ABI PRISM) on a DNA sequencer (ABI 3500XL).

Validation of CG identified MEI structure characteristics using human whole-genome PacBio data set

We analyzed WGS data based on single-molecule sequencing that was performed and publically released by Pacific Biosciences (human 54× PacBio; see URLs section) to verify specific polymorphic events that we detected by CG. The average PacBio read length was 7 680 bp. PacBio candidate insertions were locally aligned to the L1Hs sequence using the lalign program (URLs section). To find the exact insertion location in the genome we aligned each PacBio L1Hs candidate insertion junction to the human genome using the UCSC BLAT tool (GRCh37/hg19UCSC BLAT tool; see URLs section).

The same approach enabled the verification of the structure characteristics of nested and 3′ end truncated L1Hs insertions. PacBio candidate insertions were locally aligned to the L1Hs canonical sequence using a local alignment program. Candidate events that matched the canonical L1Hs sequence between nucleotides 1 to 5 500 with putative junctions in the genomic reference sequence that intersected with a pre-existing L1 repeat (hg19 RM) were referred to as nested and/or 3′-truncated L1Hs insertions.

Alignment of PacBio reads for the detection of L1Hs insertion events: The fastq format of the human 54× Pacific Biosciences dataset was downloaded from the Pacific Biosciences web site (see URLs section). The dataset included 828 fastq files which contained altogether ~22 million reads, summing up to approximately 168 billion bp. Read length statistics: Average read length was 7 680 bp, 50% of the reads were 10 739 bp long and 5% of the reads were ≥ 19 060 bp long. To align this large number of long reads we performed data parallelization as well as the following procedure: Using the Burrows-Wheeler Alignment Tool (bwasw) [70] we aligned the first and last 1 000 bp of each read (hereafter, prefix read and suffix read, respectively) to the hg19 reference. An alignment was considered valid if both prefix and suffix reads were aligned to the same chromosome and the distance between their mappings was no more than 100 000 bp. A valid aligned read was considered an insertion candidate if the following condition holds true: read length − [prefix start + suffix end + insertions − deletions] > 0, where prefix start is the start location to which the prefix read was aligned and suffix end is the end location to which the suffix read was aligned. In order to validate CG L1Hs insertion events we first selected reads containing L1Hs sequences by locally aligning PacBio candidate insertions and the L1Hs sequence using lalign (see URLs section) [71]. PacBio insertion candidates that produced significant alignments to L1Hs were subjected to further downstream analysis. In order to find the exact location in the genome into which the L1Hs was inserted, we aligned each PacBio L1Hs candidate insertion to the human genome using the UCSC BLAT tool (see URLs section).

Motif detection

L1Hs and AluY nested and non-nested insertions in the 20 brain samples and in the 78 normal LB samples were used for insertion range sequence analysis. Events with insertion ranges < 50 bp were selected, this range was extended by 10 bases in both 3′ and 5′ direction and ranges < 50 bp were analyzed for motif detection. Motif detection was done using HOMER (URLs section). We searched 4-6 base long motifs and allowed motif detection on both DNA strands. As control sets we used shuffled sequences with the same overall base composition. HOMER screens the motifs found in the target list against the background sequences for enrichment, returning motifs enriched with a $P < 0.05$.

Statistical analysis

Statistical significance was tested by either Student's *t*-test or Mann-Whitney rank-sum test for paired samples when a non-parametric statistical hypothesis test was applied on data that could not be assumed to be normally distributed. Similarity measures were tested by Pearson (linear) correlations. All reported *P* values are two-tailed unless stated otherwise. Chi-squared test (χ^2 test) was applied for the evaluation of the enrichment for the consensus sequences in L1 family elements in the genome. $P < 0.05$ indicates statistical significance.

URLs

Complete Genomics public genome data:
http://www.completegenomics.com/documents/PublicGenomes.pdf
Detailed annotation of repeats (Repeat Masker)
http://www.repeatmasker.org/
The Pacific Biosciences web site:
http://datasets.pacb.com/2014/Human54x/fastq.html
CG Public Genome Data Repository:
http://www.completegenomics.com/public-data/
CG Public Genome Data download from:
https://www.sheba-cancer.org.il/completegenomicsPublicDataMEI/
CG Public Genome Cancer Sequencing Service Guide:
http://www.completegenomics.com/documents/Cancer_Sequencing_Service_Getting_Started_Guide_2.4-2.5.pdf
Complete Genomics analysis tools; cgatools:
http://cgatools.sourceforge.net/
CG MEI detection:
http://www.completegenomics.com/FAQs/MEI-Detection/
UCSC BLAT tool (UCSC on-line version of BLAT with the Feb. 2009 (GRCh37/hg19)
https://genome.ucsc.edu/cgi-bin/hgBlat?command = start
In-house "Analysis and visualization support for Complete Genomics MEIs":
http://www.sheba-cancer.org.il/l1/
Software for motif discovery:
http://homer.salk.edu/homer/motif/
Software for multiple matching sub segments in two sequences
http://embnet.vital-it.ch/software/LALIGN_form.html

Acknowledgments

We thank the Kahn Family Foundation for their continuous support. This work was supported in part by grants from the Flight Attendant Medical Research Institute (FAMRI), and by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (Grant numbers 41/11 and 1796/12). GR is a member of the Sagol Neuroscience Network and holds the Djerassi Chair in Oncology at the Sackler Faculty of Medicine, Tel Aviv University. Human brain tissues were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD. This work was performed in partial fulfilment of the requirements for a PhD degree to JJ-H and BK, The Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University.

Author Contributions

JJ-H and GR conceived and designed the experiments. JJ-H, KC, CG, SF-B, OB and AJS performed the experiments. JJ-H, EL, RT, EE, BK, VK and MG performed the bioinformatic analysis. JJ-H, GR, NA, BK, EL and SM-M analyzed and interpreted results, and wrote the paper. SC, JR, MY, contributed to the surgical and clinical aspects of the study.

Competing Financial Interests

RT was an employee of Complete Genomics Inc. The other authors declare no competing financial interest. Whole-genome sequencing was performed by CG as an award received by GR in a competition held by CG.

References

1 Burns KHH, Boeke JDD. Human transposon tectonics. *Cell* 2012; **149**:740-752.
2 Lander ES, Linton LM, Birren B, *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**:860-921.
3 Ewing AD, Kazazian Jr HH, Kazazian HH. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* 2010; **20**:1262-1270.
4 Hancks DC, Kazazian HH. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* 2012; **22**:191-203.
5 Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* 2011; **12**:615-627.
6 Brouha B, Schustak J, Badge RM, *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* 2003; **100**:5280-5285.
7 Konkel MK, Walker JA, Hotard AB, *et al.* Sequence analysis and characterization of active human alu subfamilies based on the 1000 Genomes Pilot Project. *Genome Biol Evol* 2015; **7**:2608-2622.
8 Rechavi G, Givol D, Canaani E. Activation of a cellular oncogene by DNA rearrangement: possible involvement of an IS-like element. *Nature* 1982; **300**:607-611.
9 Amariglio EN, Hakim I, Brok-Simoni F, *et al.* Identity of rearranged LINE/c-MYC junction sequences specific for the canine transmissible venereal tumor. *Proc Natl Acad Sci USA* 1991; **88**:8136-8139.
10 Miki Y, Nishisho I, Horii A, *et al.* Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* 1992; **52**:643-645.
11 Lee E, Iskow R, Yang L, *et al.* Landscape of somatic retrotransposition in human cancers. *Science* 2012; **337**:967-971.
12 Helman E, Lawrence MLS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* 2014; **24**:1053-1063.
13 Erwin JA, Paquola ACM, Singer T, *et al.* L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* 2016; **19**:1583-1591.
14 Baillie JK, Barnett MW, Upton KR, *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 2011; **479**:534-537.
15 Coufal NG, Garcia-Perez JL, Peng GE, *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* 2009; **460**:1127-1131.
16 Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 2005; **435**:903-910.
17 Upton KR, Gerhardt DJ, Jesuadian JS, *et al.* Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* 2015; **161**:228-239.
18 Evrony GD, Cai X, Lee E, *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 2012; **151**:483-496.
19 Muotri AR, Marchetto MCN, Coufal NG, *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 2010; **468**:443-446.
20 Coufal NG, Garcia-Perez JL, Peng GE, *et al.* Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc Natl Acad Sci USA* 2011; **108**:20382-20387.
21 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**:185-188.
22 Savitsky K, Sfez S, Tagle DA, *et al.* The complete sequence of the coding region of the ATM gene reveals similarity to cell cycle regulators in different species. *Hum Mol Genet* 1995; **4**:2025-2032.
23 Crino PB, Nathanson KL, Henske EP. The Tuberous Sclerosis Complex. *N Engl J Med* 2006; **355**:1345-1356.
24 Ess KC, Kamp CA, Tu BP, Gutmann DH. Developmental origin of subependymal giant cell astrocytoma in tuberous sclerosis complex. *Neurology* 2005; **64**:1446-1449.
25 Zhou J, Shrikhande G, Xu J, *et al.* Tsc1 mutant neural stem/progenitor cells exhibit migration deficits and give rise to subependymal lesions in the lateral ventricle. *Genes Dev* 2011; **25**:1595-1600.
26 Lai MC, Lombardo MV, Baron-Cohen S. Autism. *Lancet* 2014; **383**:896-910.
27 Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* 2014; **15**:497-506.
28 Stewart C, Kural D, Strömberg MP, *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 2011; **7**:e1002236.
29 Coffee B, Cox HC, Kidd J, *et al.* Detection of somatic variants in peripheral blood lymphocytes using a next generation sequencing multigene pan cancer panel. *Cancer Genet* 2017; **211**:5-8.
30 Ross MT, Grafham D V, Coffey AJ, *et al.* The DNA sequence of the human X chromosome. *Nature* 2005; **434**:325-337.
31 Rogers JH. The origin and evolution of retroposons. *Int Rev Cytol* 1985; **93**:187-279.
32 Myers JS, Vincent BJ, Udall H, *et al.* A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* 2002; **71**:312-326.
33 Levy A, Schwartz S, Ast G. Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. *Nucleic Acids Res* 2009; **38**:1515-1530.
34 Giordano J, Ge Y, Gelfand Y, Abrusán G, Benson G, Warburton PE. Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput Biol* 2007; **3**:e137.
35 Graham T, Boissinot S. The genomic distribution of L1 elements: the role of insertion bias and natural selection. *J Biomed Biotechnol* 2006; **2006**:1-5.
36 Morrish TA, Gilbert N, Myers JS, *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* 2002; **31**:159-165.
37 Sen SK, Huang CT, Han K, Batzer Ma. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* 2007; **35**:3741-3751.
38 Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci USA* 1997; **94**:1872-1877.
39 Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Gar-

cia-Perez JL, Moran JV. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol Spectrum* 2015; **3**:1-40.
40 Yant SR, Zhu W, Millinoff D, Slightom JL, Goodman M, Gumucio DL. High affinity YY1 binding motifs: Identification of two core types (ACAT and CCAT) and distribution of potential binding sites within the human beta globin cluster. *Nucleic Acids Res* 1995; **23**:4353-4362.
41 Tesfazghi M, Riman S, Alexander Karen, Rizkallah R, Hurt M. The recruitment of the transcription factor yy1 to dna damage sites in human cells. *FASEB J* 2015; **29**:LB186.
42 Becker KG, Swergold GD, Ozato K, Thayer RE. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Hum Mol Genet* 1993; **2**:1697-1702.
43 Han JS, Szak ST, Boeke JD. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 2004; **429**:268-274.
44 Bacchelli E, Ceroni F, Pinto D, *et al.* A CTNNA3 compound heterozygous deletion implicates a role for α T-catenin in susceptibility to autism spectrum disorder. *J Neurodev Disord* 2014; **6**:17.
45 Lionel AC, Vaags AK, Sato D, *et al.* Rare exonic deletions implicate the synaptic organizer Gephyrin (GPHN) in risk for autism, schizophrenia and seizures. *Hum Mol Genet* 2013; **22**:2055-2066.
46 Piton A, Michaud JL, Peng H, *et al.* Mutations in the calcium-related gene IL1RAPL1 are associated with autism. *Hum Mol Genet* 2008; **17**:3965-3974.
47 Schmunk G, Gargus JJ. Channelopathy pathogenesis in autism spectrum disorders. *Front Genet* 2013; **4**:222.
48 Ben-Shalom R, Keeshen CM, Berrios KN, An JY, Sanders SJ, Bender KJ. Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. *Biol Psychiatry* 2017; **82**:224-232.
49 Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet* 2007; **8**:610-618.
50 Papavasiliou FN, Schatz DG. Somatic hypermutation of immunoglobulin genes. *Cell* 2002; **109**:S35-S44.
51 Beck CR, Garcia-Perez JL, Badge RM, Moran JV. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* 2011; **12**:187-215.
52 Thomas CA, Paquola ACMM, Muotri AR. LINE-1 retrotransposition in the nervous system. *Annu Rev Cell Dev Biol* 2012; **28**:555-573.
53 Simons C, Pheasant M, Makunin IV, Mattick JS. Transposon-free regions in mammalian genomes. *Genome Res* 2006; **16**:164-172.
54 Humphrey A, Higgins JNP, Yates JRW, Bolton PF. Monozygotic twins with tuberous sclerosis discordant for the severity of developmental deficits. *Neurology* 2004; **62**:795-798.
55 Miyake K, Yang C, Minakuchi Y, *et al.* Comparison of genomic and epigenomic expression in monozygotic twins discordant for Rett Syndrome. *PLoS One* 2013; **8**:e66729.
56 Hu VW, Frank BC, Heine S, Lee NH, Quackenbush J. Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC Genomics* 2006; **7**:118.
57 Li W, Jin Y, Prazak L, Hammell M, Dubnau J. Transposable elements in TDP-43-mediated neurodegenerative disorders. *PLoS One* 2012; **7**:1-10.
58 Bundo M, Toyoshima M, Okada Y, *et al.* Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* 2014; **81**:306-313.
59 Eickbush TH. Repair by retrotransposition. *Nat Genet* 2002; **31**:126-127.
60 Ono R, Ishii M, Fujihara Y, *et al.* Double strand break repair by capture of retrotransposon sequences and reverse-transcribed spliced mRNA sequences in mouse zygotes. *Sci Rep* 2015; **5**:12281.
61 Doucet-O'Hare TT, Rodić 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* USA 2015; 112: E4894-E4900.
62 Suberbielle E, Sanchez PE, Kravitz A V, *et al.* Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-β. *Nat Neurosci* 2013; **16**:613-621.
63 Adamec E, Vonsattel JP, Nixon RA. DNA strand breaks in Alzheimer's disease. *Brain Res* 1999; **849**:67-77.
64 Gao Y, Sun Y, Frank KM, *et al*. A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* 1998; **95**:891-902.
65 Wei PC, Chang AN, Kao J, *et al.* Long neural genes harbor recurrent dna break clusters in neural stem/progenitor cells. *Cell* 2016; **164**:644-655.
66 Drmanac R, Sparks AB, Callow MJ, *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010; **327**:78-81.
67 Carnevali P, Baccash J, Halpern AL, *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* 2012; **19**:279-92.
68 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**:841-842.
69 Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* 2010; **Chapter 11**:Unit 11.7.
70 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**:1754-1760.
71 Huang X, Miller W. A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* 1991; **12**:337-357.
72 Carmi S, Hui KY, Kochav E, *et al*. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun* 2014; **5**:4835.
73 C Yuen RK, Merico D, Bookman M, *et al*. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* 2017; **20**:602-611.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)