Cell

Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain

Gilad D. Evrony,[1,5,6,11] Xuyu Cai,[1,5,6,11] Eunjung Lee,[2,9] L. Benjamin Hills,[5,6] Princess C. Elhosary,[7] Hillel S. Lehmann,[5,6] J.J. Parker,[5,6] Kutay D. Atabay,[5,6] Edward C. Gilmore,[10] Annapurna Poduri,[3,7] Peter J. Park,[2,8,9] and Christopher A. Walsh[1,3,4,5,6,*]

[1]Program in Biological and Biomedical Sciences
[2]Center for Biomedical Informatics
[3]Department of Neurology
[4]Department of Pediatrics
Harvard Medical School, Boston, MA 02115, USA
[5]Division of Genetics, Manton Center for Orphan Disease Research
[6]Howard Hughes Medical Institute
[7]Department of Neurology
[8]Informatics Program
Boston Children's Hospital, Boston, MA 02115, USA
[9]Division of Genetics, Brigham and Women's Hospital Boston, MA 02115, USA
[10]Department of Pediatrics, Case Western Reserve University School of Medicine, Cleveland 44106, OH
[11]These authors contributed equally to this work
*Correspondence: christopher.walsh@childrens.harvard.edu
http://dx.doi.org/10.1016/j.cell.2012.09.035

SUMMARY

A major unanswered question in neuroscience is whether there exists genomic variability between individual neurons of the brain, contributing to functional diversity or to an unexplained burden of neurological disease. To address this question, we developed a method to amplify genomes of single neurons from human brains. Because recent reports suggest frequent LINE-1 (L1) retrotransposition in human brains, we performed genome-wide L1 insertion profiling of 300 single neurons from cerebral cortex and caudate nucleus of three normal individuals, recovering >80% of germline insertions from single neurons. While we find somatic L1 insertions, we estimate <0.6 unique somatic insertions per neuron, and most neurons lack detectable somatic insertions, suggesting that L1 is not a major generator of neuronal diversity in cortex and caudate. We then genotyped single cortical cells to characterize the mosaicism of a somatic *AKT3* mutation identified in a child with hemimegalencephaly. Single-neuron sequencing allows systematic assessment of genomic diversity in the human brain.

INTRODUCTION

It is unlikely that the genomes of any two cells in the body are identical, due to somatic mutations during replication and other mutagenic forces (Frumkin et al., 2005). The complexity and diversity of neuronal cell types in the brain have also led to suggestions that a somatic mutational mechanism may have been harnessed evolutionarily to diversify neuronal function (Muotri and Gage, 2006; Rehen et al., 2005). Endogenous retrotransposition of LINE-1 elements has been proposed as one potential mechanism generating neuronal genome diversity (Singer et al., 2010). Human-specific LINE-1 (L1Hs) retrotransposons comprise the only known active autonomous transposon family in humans, with ~80–100 active L1Hs elements per individual (Hancks and Kazazian, 2012), and somatic L1Hs insertions have been found both in cancerous and normal cells (Iskow et al., 2010; Lee et al., 2012a; Miki et al., 1992; Van den Hurk et al., 2007). Recent studies observed rare retrotransposition of an L1Hs reporter in rodent brain in vivo (Muotri et al., 2005, 2010) and human neural progenitors in vitro (Coufal et al., 2009), whereas other studies found evidence for more widespread somatic L1Hs insertions in the human brain by qPCR (Coufal et al., 2009) and bulk DNA sequencing (Baillie et al., 2011). qPCR estimates of these events in human brain approach 80 somatic insertions per cell (Coufal et al., 2009).

Although L1 retrotransposition and other somatic mutations could contribute to functional genomic diversity, they can also cause disease (Erickson, 2010; Hancks and Kazazian, 2012). Therefore, any potential somatic mutational mechanism must be balanced by the need for genome stability. Somatic mutations cause not only cancers, but also several malformations of the brain (Gleeson et al., 2000; Rivière et al., 2012), emphasized by the recent identification of somatic mutations affecting genes of the PI3K-AKT3-mTOR pathway in hemimegalencephaly (HMG) (Lee et al., 2012b; Poduri et al., 2012), a severe epileptic

brain malformation. However, the rates and types of somatic mutations occurring during normal brain development and how much of the unexplained burden of neurogenetic disease may be caused by somatic mutations are unknown (Erickson, 2010).

Systematically studying somatic mutations requires sequencing genomes of single cells (Kalisky et al., 2011) because the signals of somatic mutations present in a minority of cells can be missed due to sequencing error or insufficient sequencing depth. Single-cell sequencing overcomes this limitation, as shown by studies of single human cancer cells and single sperm that have yielded important new insights into tumor evolution and genetic heterogeneity (Hou et al., 2012; Navin et al., 2011; Wang et al., 2012; Xu et al., 2012). However, similar technologies have yet to be applied to the study of somatic mutation in normal human tissues such as brain or to diseases other than cancer.

Here, we describe a method to amplify genomes of single neurons from postmortem and surgically resected human brain, enabling interrogation of a wide range of somatic mutations by high-throughput sequencing. We performed genome-wide L1Hs insertion profiling of 300 single neurons from cerebral cortex and caudate nucleus of three neurologically normal individuals and confirmed that somatic L1Hs retrotransposon insertions are present in the normal human brain. Our quantitative analysis of >200,000 L1Hs insertion sites in these 300 single neurons suggests a frequency not higher than 0.6 unique somatic insertions per neuron and possibly as low as 0.04 (1 insertion in 25 neurons), consistent with observed in vitro rates for human neural progenitors but substantially less than previous qPCR-based estimates for human brain (Coufal et al., 2009). We then sequenced single cells from HMG brain tissue harboring a known somatic *AKT3* point mutation (c.49G→A; p.E17K) (Poduri et al., 2012), showing that our method can characterize the mosaicism of pathogenic somatic brain mutations. These single-cell studies provide a foundation for studying genomic variability among cells in the human brain, both in normal development and in neurologic disease.

RESULTS

High-Throughput Isolation and Amplification of Single Neuronal Genomes from Human Brains

We purified nuclei from postmortem human frontal cortex and caudate nucleus and labeled them with a neuron-specific antibody (NeuN) for sorting using fluorescence-activated cell sorting (FACS) (Figure 1A) (Matevossian and Akbarian, 2008; Spalding et al., 2005). Large nuclei with neuronal nuclear morphology (Parent and Carpenter, 1996) were readily apparent by microscopy (Figure S1A). NeuN immunoreactivity (Figure S1B) (Mullen et al., 1992) labels essentially all neuronal nuclei in cortex and caudate (Wolf et al., 1996), corresponding to 25%–35% of all nuclei (population I; Figures 1B and S1C). Consistent with their increased size on microscopy (Figure S1B), $NeuN^+$ nuclei also had larger forward (FSC) and side (SSC) scatter (correlates of size) by flow cytometry compared to $NeuN^-$ nuclei (Figure S1D). Whereas for nuclei isolated from the caudate we performed a simple sort of the $NeuN^+$ population (population I; Figure S1C), we further enriched nuclei from the cortex for pyramidal neuronal nuclei. Because neighboring cortical pyramidal neurons tend to have shared clonal origins due to their primarily radial migration (Magavi et al., 2012), enriching for pyramidal neuronal nuclei increases the chance of identifying clonal somatic mutations shared by multiple neurons. The largest neuronal nuclei in cortex correspond primarily to pyramidal projection neurons (Gittins and Harrison, 2004; Mills, 2007), and indeed, their nuclei often show a pyramidal shape (Figure S1A). We therefore sorted cortical nuclei within the top 25% NeuN/FL-2 fluorescence of population I (population Ia; Figure 1B), which were the largest nuclei in population I (Figure S1D). We confirmed the neuronal and nonneuronal identities of the sorted populations by reverse transcriptase PCR (RT-PCR) and western blot analysis of additional neuronal (*SNAP25* and *SYT1*) and nonneuronal (*GFAP*, *AQP4*, and Olig2) markers (Figures 1C and 1D). For every sort, a portion of the sorted nuclei was reanalyzed by FACS, confirming that nuclei remained intact during sorting and that sort purity was >98% (Figures 1B and S1C).

We used multiple displacement amplification (MDA) (Dean et al., 2002) for whole-genome amplification of single nuclei because it produces large yields of high molecular weight amplicons, most of which are >30 kb (Hou et al., 2012 and data not shown), allowing study of both single-nucleotide mutations and ~6 kb full-length L1Hs insertions. We optimized MDA reactions for increased yield (Figure S1E), producing 15–20 μg of amplified DNA from single cells. We also measured exogenous (non-human) DNA contamination in the reagents of the MDA reaction (Blainey and Quake, 2011), finding negligible (<1 fg) exogenous DNA (Figures S1F and S1G). Additional controls (see following section) excluded operator human DNA contamination. Quantitative MDA (qMDA) reactions (Zhang et al., 2006) further showed that, as the number of nuclei sorted in a well increased, the time-to-threshold-amplification decreased in a stepwise manner ($p < 0.01$ for each additional nucleus) (Figure 1E), confirming that the desired number of nuclei was correctly sorted in each well. We concluded that our procedure can sort and amplify single neuronal genomes from human brains with high purity and in a high-throughput manner.

Genome-wide Coverage and Amplification Dropout Rates of Single Neuronal Genomes

We next evaluated the genome-wide coverage and reproducibility of our single neuronal genome amplification. In an initial four-locus multiplex PCR quality control, 97% of sorted single neurons amplified at least three of the four loci, indicating that their genomes were successfully amplified and suitable for further experiments. We then performed low-coverage whole-genome sequencing (Figure 2A) of eight randomly chosen single neurons (0.35× average coverage)—six from a normal individual (46XY) and two from a trisomy 18 individual—as well as unamplified and MDA-amplified bulk reference samples. The two neurons from the trisomy 18 individual showed the expected increase in chromosome 18 copy number, and the six single neurons from the normal individual were all euploid, confirming that intact nuclei were sorted and that all chromosomes were amplified (Figure 2B). Counting sequencing reads across the genome in bins ~500 kb in size (Navin et al., 2011) revealed

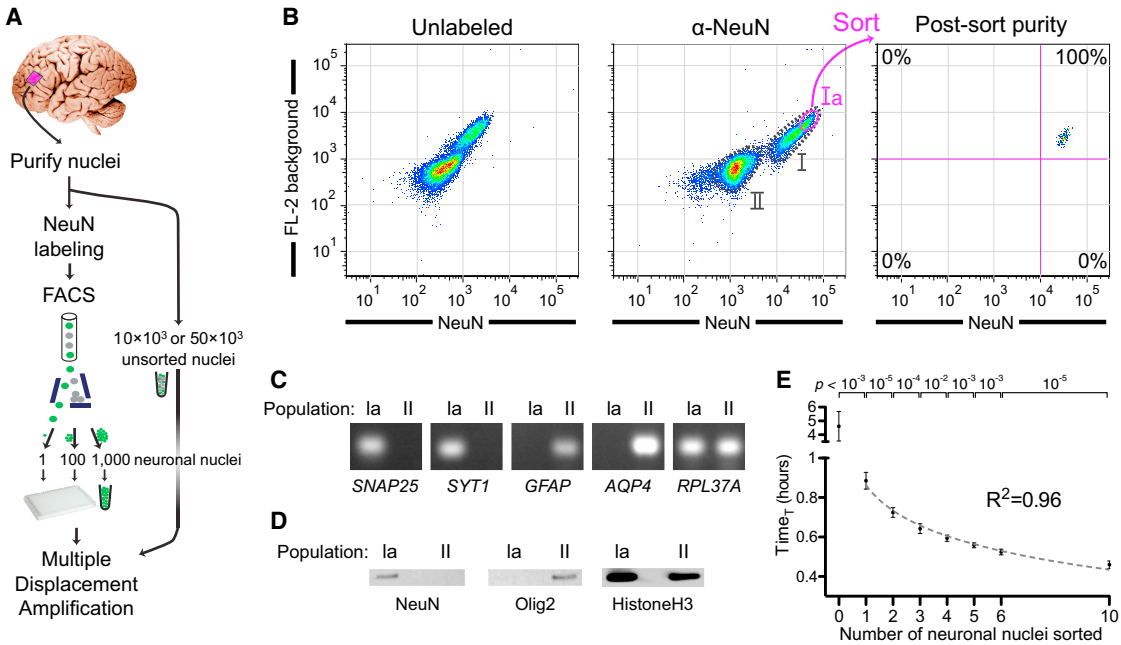

Figure 1. Isolation and Genome Amplification of Single Human Neuronal Nuclei

(A) Schematic of the method.

(B) Fluorescence-activated cell sorting of cortical nuclei stained with NeuN shows two separable populations: NeuN$^+$ (population I) and NeuN$^-$ (population II). A subset of population I (Ia) consisting of large neuronal nuclei was sorted and reanalyzed, confirming sort purity. Two populations of nuclei are sometimes apparent without NeuN staining due to the increased background staining of the larger population I nuclei. Fluorescence decrease of the sorted population on re-analysis is always observed due to photobleaching and washing of nonspecific staining in the first sort.

(C) RT-PCR confirming the neuronal and nonneuronal identities of populations Ia and II, respectively, by assaying for expression of nuclear RNA for two neuronal (*SNAP25* and *SYT1*), two astroglial (*GFAP* and *AQP4*), and input control (*RPL37A*) genes. RT-PCR and western blot experiments (Figures 1C and 1D) were performed with NeuN/Mef2c double labeling in which all NeuN$^+$ nuclei were Mef2c$^+$ (data not shown).

(D) Western blot analysis of NeuN and Olig2 (an oligodendrocyte marker), confirming neuronal and nonneuronal identity, respectively, of populations Ia and II.

(E) Quantitative MDA reactions monitored in real-time confirm accurate sorting of the desired number of nuclei. The time to amplify to a threshold above background ($Time_T$, analogous to qPCR C_T value) is plotted on the y axis (error bars ± 1 SD; n = 7 or 8 reactions per condition). Points were fit to a semi-log line of slope −4.3, corresponding to 1.7-fold amplification per unit time.

See also Figure S1, and see Table S3 for RT-PCR primer sequences.

a systematic, regional amplification bias for all MDA samples, compared to unamplified bulk DNA, regardless of the number of nuclei amplified (Figure S2A). This regional bias in MDA amplification could be controlled for using any of the MDA samples as a reference (Figure 2C), indicating that most of the regional variability in amplification is inherent to MDA rather than the number of nuclei amplified. Bias in amplification relative to GC content was also similar for all MDA samples types (Figure S2B).

In order to use single-neuron sequencing for somatic mutation detection, amplified genomes must reflect the diploid genotype (both alleles) of genomic loci. We therefore quantified the fraction of genomic loci that failed to amplify one (allelic dropout, AD) or both alleles (locus dropout, LD). Loss of one allele, AD, was measured with a panel of 16 polymorphic microsatellite markers (Identifiler fingerprinting) and by SNP microarray genotyping. AD measured by Identifiler of 92 single neurons across 1,183 heterozygous loci was 9.5% (Figure 2D), whereas AD measured by SNP microarray (for >60,000 loci that are heterozygous in the bulk DNA and called with high confidence in both the reference and sample) was 8%–9% in three single neurons (Figure S2C and Table S1A), consistent with previous estimates (Hou et al., 2012). Some dropout tended to recur at specific loci even in MDA-amplified 100- and 1,000-neuron samples (Figure S2D), probably reflecting difficulty of MDA to amplify specific loci. Loss of both alleles, LD (locus dropout), was 2.3% in the 92 single neurons assayed by Identifiler. In addition, LD was separately estimated by counting the percentage of low-coverage sequencing bins with less than 1/16 the copy number relative to an unamplified DNA reference and was 2.0% for 1-neuron samples (Figure S2E). These low rates of AD (~10%) and LD (~2%) demonstrate comprehensive and reproducible amplification of single neuronal genomes and suggest that genome-wide profiling of L1 insertions in single neurons could capture up to 90% of retrotransposon insertions per cell. These genotyping controls also excluded operator contamination, as all amplified single neuronal genomes tested were concordant with the bulk reference (Figures 2D and 2E and Tables S1B–S1C).

Genome-wide L1Hs Profiling in Single Neurons

We performed genome-wide L1Hs insertion profiling (L1-IP) of single neurons by adapting the method of Ewing and Kazazian (2010) for high-throughput multiplexed sequencing. All known

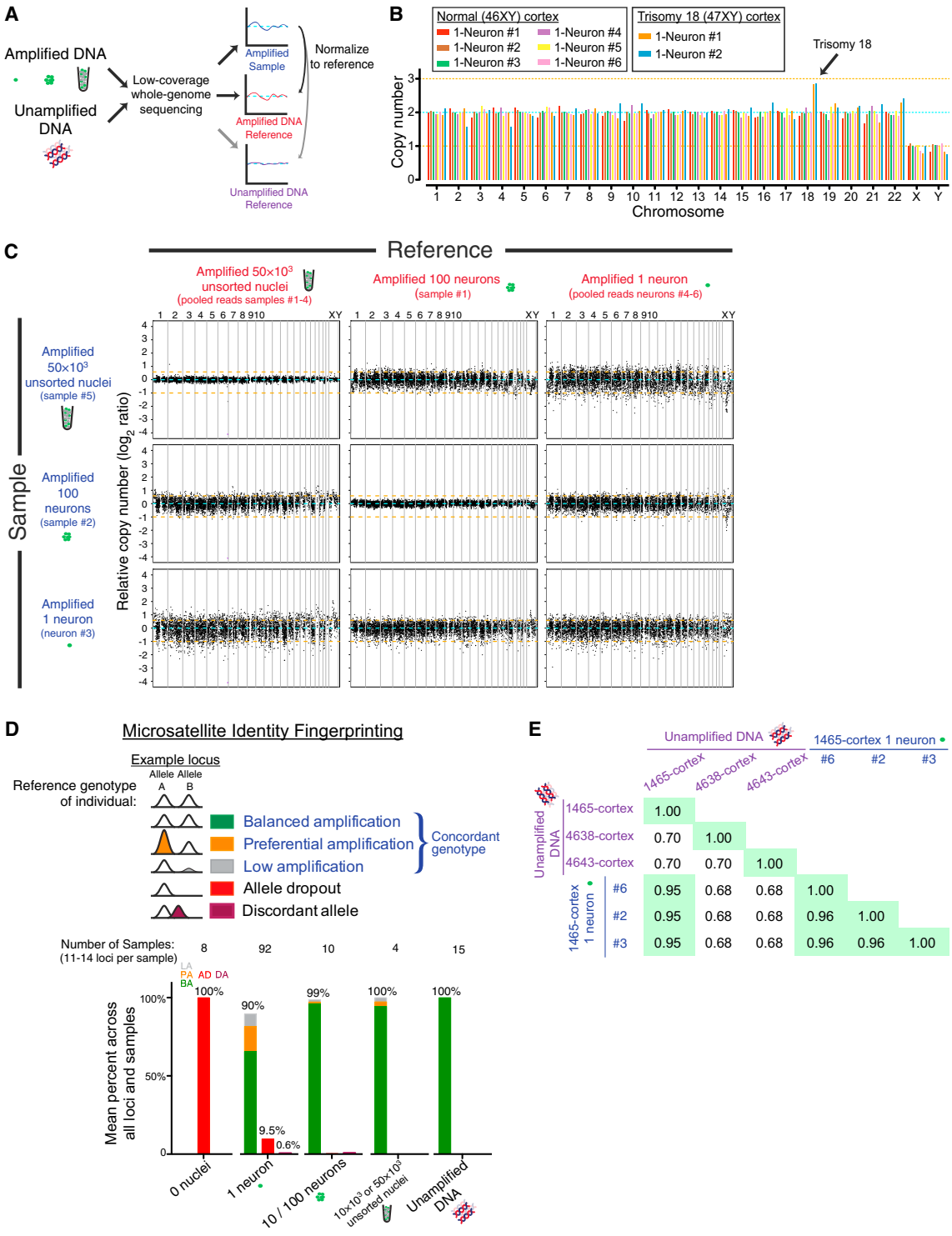

Figure 2. Single-Neuron Genome-wide Coverage, Amplification Bias, and Identity Fingerprinting

(A) Schematic of the low-coverage whole-genome sequencing method.

(B) Chromosome copy numbers of single cortical neurons from normal (UMB1465, 46XY) and trisomy 18 (UMB866, 47XY, +18) individuals. Copy numbers are normalized to the median copy number of each chromosome across the eight single neurons, with autosomes adjusted to a median copy number of 2. Orange lines denote ± 1 copy.

(C) Higher-resolution copy number profiling in 6,000 equal-read bins of ~500 kb in size shows that MDA bias can be corrected by normalization to an MDA-amplified reference. Orange lines denote ± 1 copy, and purple points indicate off-scale bins.

active and disease-causing L1Hs subfamilies possess two sequences diagnostic of L1Hs (Hancks and Kazazian, 2012; Ovchinnikov et al., 2002), and a comprehensive study of somatic insertions in the setting of cancer found that 110/111 somatic insertions (with evidence of a target site duplication and poly-A tail) contained both sequences (Lee et al., 2012a). L1-IP targets these L1Hs-specific sequences and amplifies genomic DNA flanking L1Hs insertions containing these diagnostic sequences (Figures 3A, 3B, and S3A).

We profiled from each of three neurologically normal individuals: 50 single neurons from cerebral cortex and 50 from caudate nucleus (i.e., 300 MDA-amplified single neurons total); unamplified bulk DNA from 5–6 tissues (cortex, caudate, cerebellum, heart, liver, and lung); MDA-amplified 50,000-cell, 10,000-cell, 1,000-cell, and 100-neuron samples; and technical replicates to assess reproducibility (Figures S3B and S3C), for a total of 383 samples (see Table S2 for sample details). A custom data analysis pipeline classified detected peaks as known reference insertions present in the human genome reference (KR), known nonreference insertions identified in previous studies (KNR), or unknown (UNK) candidate insertions and assigned a confidence score ranging from 0 to 1 (low-quality to high-quality peaks) based on the number of reads and the number of unique read start sites per peak (Figure 3A). The confidence score was derived from a logistic regression model of germline insertions reproducibly found in bulk DNA samples of the individual (Figure S3D and see Extended Experimental Procedures for details of the analysis pipeline).

MDA is known to produce rare, low-level chimeric sequences due to local, occasional mispriming of single-stranded amplicons to each other during amplification (Lasken and Stockwell, 2007). These chimeras were seen in MDA-amplified samples as an excess of background reads and peaks with low read depth and one or few unique read start sites in the local ~20 kb flanks of some though not all L1 insertions (Figures 3B and S4A–S4D). Because chimeras form at different sites in different MDA reactions, they are not recurrent between samples (Figures S5A and S5B), and cloning of chimeras (representative example in Figures S5A–S5C) confirmed their MDA-derived mechanism of formation. Their low confidence scores (Figure S4B) allowed most MDA-chimera peaks to be filtered with minimal reduction in sensitivity for known insertions (Figure 3C).

We first assessed the sensitivity of L1-IP to detect L1Hs insertions genome-wide. In 1-neuron samples, the sensitivity of L1-IP for KR insertions (mostly homozygous) present in bulk DNA of the individual was 81% ± 6% (SD), with a confidence score threshold of 0.5 (Figure S6A), and of 300 1-neuron samples in this study, only four were low-quality outliers (Figure S6B). Sensitivity increased to 87% when relaxing the confidence threshold to 0.1, though at this lower confidence score, more candidate insertions with weaker evidence supporting them were also detected. Because somatic insertions are expected to be present in a single copy, sensitivity for single-copy insertions in 1-neuron samples was assessed with chrX KR/KNR insertions in individual 1465 (male) and was only slightly lower at 75% ± 10%, with a confidence score threshold of 0.5 (Figure S6A). We further confirmed that we detect the expected absolute number of insertions: the mean number of KR, KNR, and UNK insertions (with confidence score >0.5) per bulk DNA sample was 689, 113, and 43, respectively (Figure S6C), compared to 628 KR and 152 KNR/UNK insertions found on average in a previous study (Ewing and Kazazian, 2010). 605, 87, and 47 KR, KNR, and UNK peaks were found on average in 1-neuron samples (Figure S6C). A plot of L1Hs peaks found in bulk DNA, a 100-neuron sample, and two representative single neurons is shown in Figure 4.

In order to validate L1-IP-predicted insertions, we optimized a 3′ junction PCR validation method (3′PCR) (Figure S6D) and further used it to directly measure allelic dropout (AD) and locus dropout (LD) of L1Hs insertions in amplified single neurons. The technical sensitivity of the 3′PCR validation method (i.e., 3′PCR detection rate of true germline insertions) was important to determine first in order to estimate at what rate true insertions found by L1-IP fail to validate by 3′PCR. This was assayed by 3′PCR of 64 known germline insertions (33 KR and 31 KNR) in unamplified bulk DNA and amplified unsorted 50,000-nuclei and 1-neuron samples. In 1-neuron samples, 3′PCR detected 94% of known germline insertions with the first primer attempted (the remainder were validated successfully with redesigned primers), and this detection rate was not significantly different between amplified and unamplified samples (Figures 3D and S6E). 3′PCR can therefore sensitively detect L1Hs insertions in amplified single neuronal genomes. 3′PCR also successfully validated, in both bulk and 1-neuron samples, 12 out of 12 unknown (UNK) germline candidate insertions that we tested (Figures 3D, S6E, and Table S3), confirming that L1-IP can identify unknown germline insertions. AD of L1Hs insertions was then estimated by 3′PCR of three heterozygous insertions in a larger number of 83 single neurons (Figures 3E, S6F, and S6G), finding 8.0% AD (20/249 alleles), consistent with previous estimates. LD estimated by 3′PCR of three homozygous insertions in the same cells (Figures 3E and S6G) was 1.2% (3/249 alleles). We concluded that L1-IP's high sensitivity to detect germline insertions in single neurons, our robust 3′PCR validation method, and direct confirmation of <10% L1Hs allelic dropout allow us to confidently search for somatic L1Hs insertions genome-wide in single neurons.

Identity Fingerprinting of Single Neurons by L1Hs Profile

L1-IP can reliably detect population-polymorphic L1Hs insertions in single neurons (Figures 5A–5C), serving as a fingerprint for each individual. All possible permutations of insertion polymorphisms among the three individuals were found (every possible pair of individuals and individual specific), and as expected, KR and KNR insertions were enriched in fixed and

(D) Identifiler fingerprinting confirms that the single neurons derive from the correct individuals and measures allele preferential amplification (PA), low amplification (LA), allele dropout (AD), and discordant allele (DA) rates.

(E) Fraction of genotypes by SNP microarray that are concordant between three single neurons and bulk DNA confirms that the single neurons derive from the correct individual.

See also Figure S2 and Table S1.

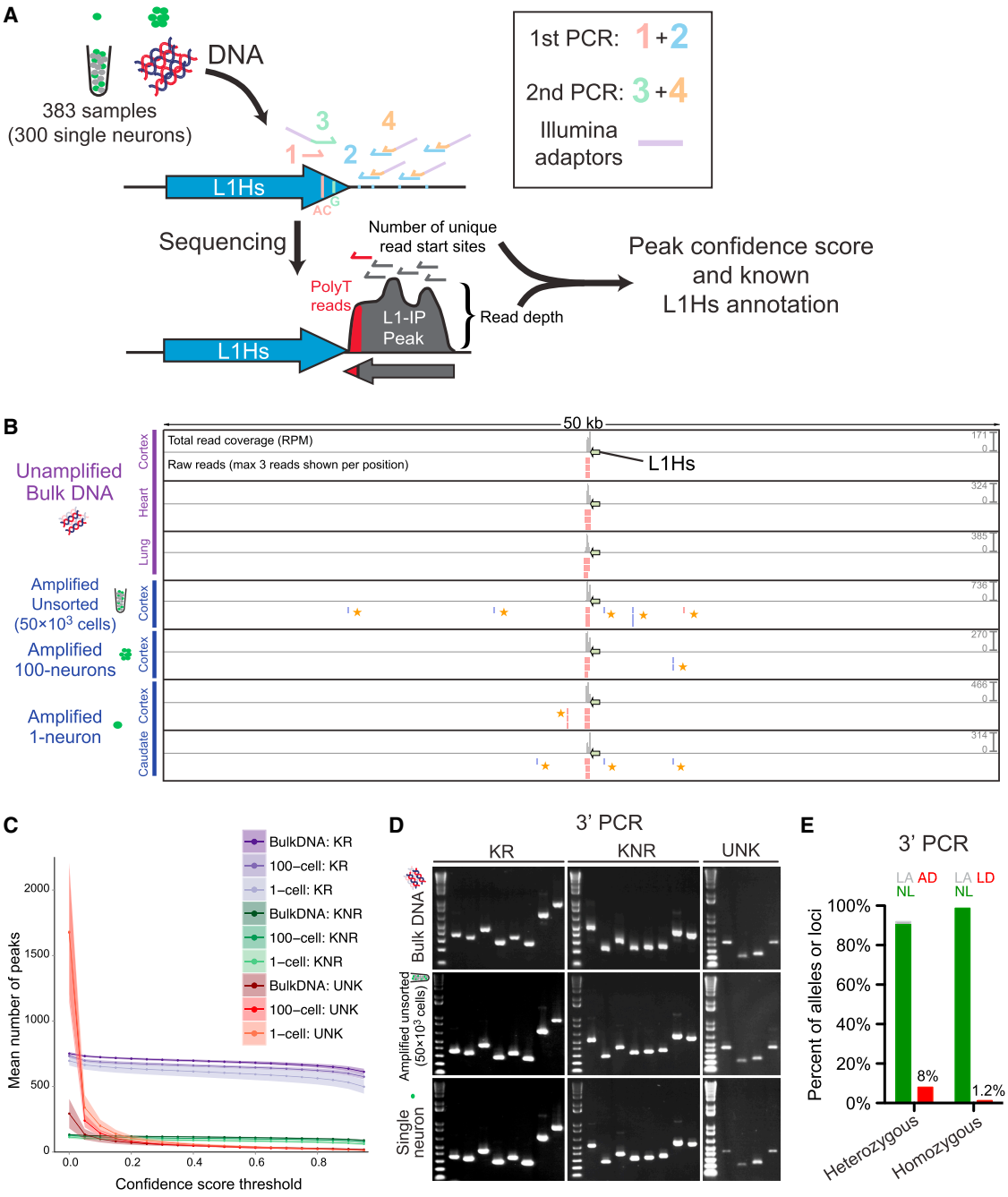

Figure 3. Genome-wide L1Hs Insertion Profiling in Single Neurons

(A) Schematic of the L1-IP method. Primers 1 and 3 (L1Hs-AC and ILMN-Adaptor1_L1Hs-G, respectively) are specific to L1Hs diagnostic nucleotides ("AC" and "G"). Primer 2 represents eight different 5 bp arbitrary seed primers, each containing the same barcode. Primer 4 (ILMN-SeqAdaptor2) incorporates an Illumina adaptor. See Table S3 for primer sequences.

(B) L1-IP sequencing reads for one representative known reference insertion (L1Hs-KR-chr11_115209613). For each sample, a total read coverage track and a raw reads track are shown. Each read coverage track is scaled to the maximum peak height of the sample (scale on the right, in reads per million mapped reads [RPM]). In the raw reads track, up to three reads are shown for each position. The green arrow marks the L1Hs insertion. Plus and minus strand reads are red and blue, respectively. Low-level MDA-chimera reads (yellow asterisks) are seen in the local region of the true insertion only in MDA-amplified samples.

(C) The number of peaks found above different confidence score thresholds corresponding to known reference insertions (KR), known nonreference insertions (KNR), and unknown peaks (UNK). Data shown are the mean for all bulk (n = 31), 100-cell (n = 15), and 1-cell (n = 303) samples from all three individuals (includes 15, 5, and 3 technical replicates, respectively). Shading around each line shows ± SD. KR and KNR insertions used for peak annotation are in Table S5.

polymorphic insertions, respectively (Figure 5A). Hierarchical clustering of all samples in the study according to L1Hs genotype correctly clustered all samples by individual except for three low-quality 1-neuron samples (Figure 5A). Importantly, because both population-polymorphic and somatic insertions belong to the same L1Hs subfamilies and have the same L1Hs diagnostic nucleotides (Beck et al., 2010; Lee et al., 2012a), detection of population-polymorphic L1Hs insertions in single neuronal genomes further illustrates that L1-IP has the potential to capture somatic insertions.

Somatic L1Hs Insertion Rate in Cortex and Caudate Neurons

Our single-neuron L1-IP data allowed us to quantify the number of cortex- and caudate-specific somatic insertions in single-neuron samples and estimate an upper bound for the number of somatic L1Hs insertions per neuron (defined as absent from bulk DNA samples of the individual excluding the brain region being analyzed). Rather than using the same confidence score threshold across all samples, we adjusted the confidence score threshold for each single-neuron sample to maintain a constant sensitivity for KNR germline insertions. This controls for variability in single-neuron sample quality and allows for more accurate correction of insertion rates for sensitivity. A KNR reference was specifically chosen, as it would be expected to better estimate sensitivity for single-copy somatic events than a mostly homozygous KR reference set. We excluded insertions found within 20 kb of known (KR/KNR) insertions, leading to a minimal reduction in sensitivity (by excluding 1.5% of the genome, i.e., 45.5/3137 Mb) with a substantial gain in specificity by filtering most though not all MDA chimera peaks (Figure S4A). At a sensitivity threshold that detects 50% of KNR insertions, we found an average of 1.1 ± 2.3 (SD) somatic insertion candidates per neuron (corrected for sensitivity) (Figure 6A), and 68% of 1-neuron samples had no detectable somatic insertions. Additionally, we counted the number of unique somatic insertions per neuron (i.e., not present in other single neurons sequenced from the individual) and found 0.6 ± 1.5 (SD) candidate unique insertions per neuron (Figure 6B); 82% of 1-neuron samples had no detectable unique somatic insertions.

The above upper bound estimate for the somatic insertion rate controls for sensitivity (i.e., false negative rate) but is likely an overestimate, as it does not take into account specificity (i.e., false positive MDA chimera and other artifactual peaks still remaining after our sensitivity threshold and local 20 kb filtering). We therefore screened for false positive candidates by carrying out 3′PCR validation and secondary validations of the 16 highest-scoring candidate somatic insertions from each tissue (96 total). Initial review of L1-IP raw data revealed that at least half of the candidates were likely MDA-chimeras or other recognizable technical artifacts that cannot be systematically filtered. These include peaks caused by read alignment errors, chimeras of older L1Pa insertions, and loci with systematic low-level reads present at subthreshold levels in many unamplified bulk and MDA-amplified samples of unrelated individuals but stochastically passing threshold as somatic candidates in one or a few single-neuron samples (see Table S3 for annotation of the 96 candidates). Indeed, only 17 of the 81 candidates (21%) for which we could design primers passed 3′PCR validation (Figure S7A), significantly less than the 94% validation rate for known insertions (Figure S6E). Secondary validation sequencing of 3′PCR products and review of L1-IP raw data revealed that 12 of the remaining 17 candidates were chimeras or nonspecific PCR products. Therefore, most of the somatic candidates are likely false positives, and the true somatic L1Hs insertion rate may be significantly lower than our upper bound estimate prior to validation. The postvalidation somatic and unique somatic insertion rate estimates are 0.07 ± 0.15 (SD) and 0.04 ± 0.10 (SD) insertions per neuron, respectively (Figures 6A and 6B).

The remaining five somatic candidates were studied further by attempting to clone their full lengths and screening for their presence by 3′PCR across all single neurons sorted from the individual in which they were found. We successfully cloned the full length of one of the five somatic insertion candidates (Figure S7B). This insertion was detected in our L1-IP data in intron 4 of the gene *IQCH* (IQ motif containing H, chromosome 15), in neuron #2 from the cortex of individual 1465, and is a full-length, intact 6.1 kb L1Hs with all of the hallmarks of a bona fide L1Hs insertion: a target site duplication (TSD) (13 bp), a poly-A tail (~71 bp), and a 5′ transduction (101 bp), allowing us to trace its source to a full-length, population-polymorphic KR L1Hs on chromosome 8 (Figures 6D, S7C and S7D). The full-length sequence of the somatic insertion (Table S3) precisely matched the sequence of the source L1Hs. The insertion was not detected by standard 3′PCR in brain and nonbrain bulk tissues from the individual (Figure 6C) and was found in 2/83 (2.4%) cortical and 0/59 caudate single neurons tested (Figure 6E). The insertion was detected at low levels in L1-IP data of some unsorted 50,000-nuclei samples (Figure S7E), as expected for a low-level mosaic insertion, and with further optimization of our 3′PCR protocol (increased DNA input and higher-cycle PCR), we were able to amplify the insertion from these bulk samples as well (Figure S7F). The remaining four candidates were each found by 3′PCR only in the single neuron in which they were identified by L1-IP. Three of the four had poly-A tails by 3′PCR product sequencing (the fourth had an indeterminate poly-A tail because the breakpoint was within a genomic poly-A) (Table S3). Our results illustrate the ability of single-cell sequencing to identify somatic L1Hs insertions and highlight the potential of single-cell sequencing to identify very low-level mosaic mutations in human tissue.

(D) Representative gel images of 3′ junction PCR (3′PCR) of 20 different germline insertions (8 KR, 8 KNR, and 4 UNK).
(E) 3′PCR quantification of AD and LD in 1-neuron samples (n = 83) of three heterozygous and three homozygous L1Hs insertions. AD and LD are quantified for heterozygous and homozygous insertions, respectively. NL, normal amplification; LA, low amplification; AD, allelic dropout; LD, locus dropout.
See also Figures S3, S4, S5, and S6, and Tables S2, S3, and S5.

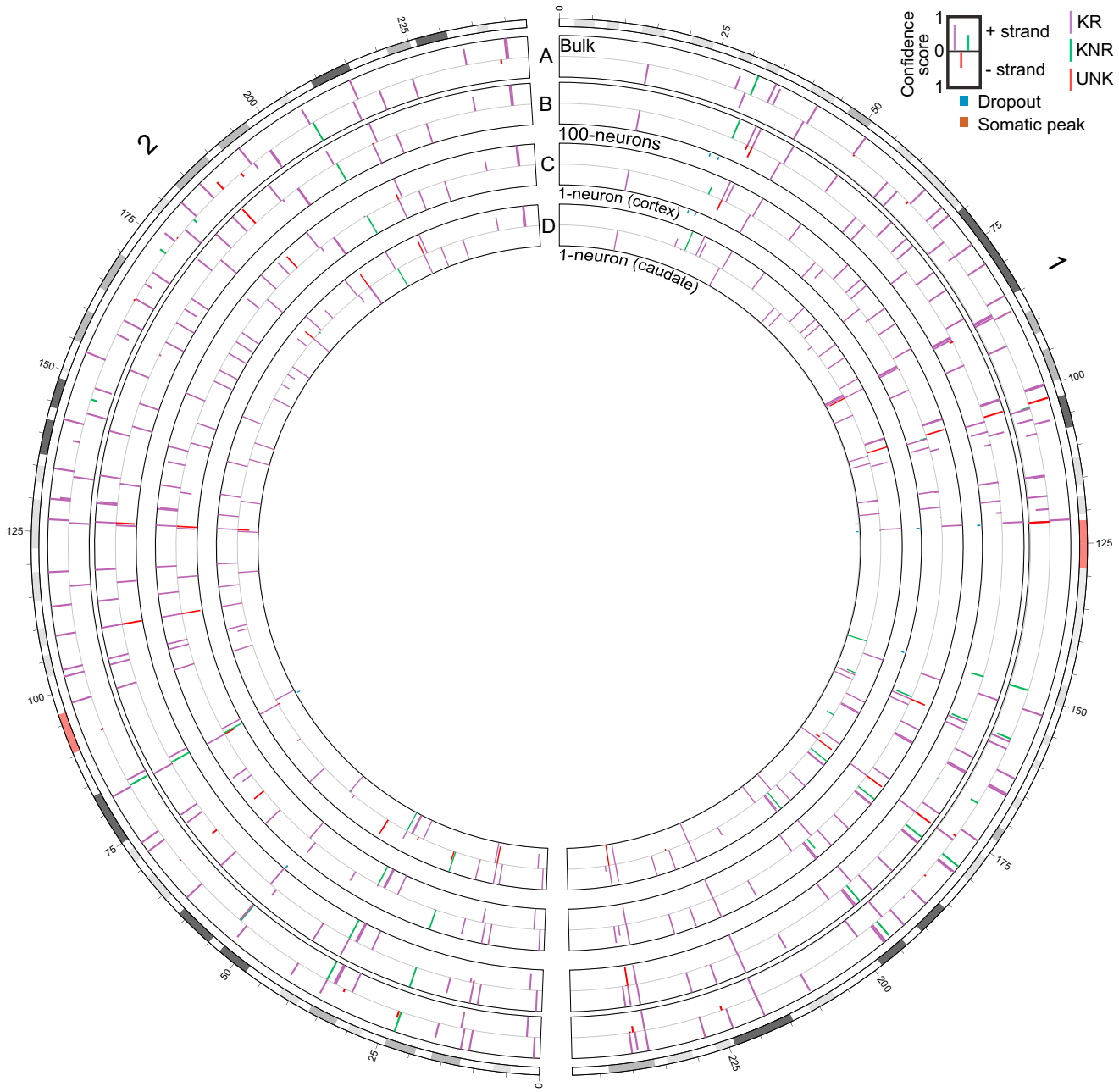

Figure 4. Chromosome L1-IP Profile of Single Neurons

(A–D) Circos plot (Krzywinski et al., 2009) of chromosomes 1 and 2 from representative L1-IP samples from individual 1465: (A) bulk DNA, (B) cortex 100-neurons #1, (C) cortex 1-neuron #2, and (D) caudate 1-neuron #1. Peaks are shown for loci in which at least one of the samples has a peak confidence score > 0.5. Bulk DNA track shows the mean confidence score across all bulk DNA samples of individual 1465. KR, KNR, and UNK peaks are colored as indicated in the key. Below 100-neuron and 1-neuron sample tracks are annotations for peaks present with a score >0.5 in bulk DNA but absent in the sample ("Dropout") and peaks absent from bulk DNA but present in the sample with a score >0.5 and at least 20 kb away from the nearest KR/KNR insertion in the individual to exclude MDA-chimera peaks ("Somatic peak"). Figures for all chromosomes can be found in Data S1.

Single-Cell Sequencing Quantifies Mosaicism of a Somatic Brain Mutation Causing Hemimegalencephaly

Given the low rate of L1 retrotransposition in neocortical progenitors of normal brains, we next studied the ability of single-neuron sequencing to characterize a pathogenic somatic point mutation in the brain. An open question regarding the pathophysiology of hemimegalencephaly is the lineage (developmental origin) of the pathologic cells (Flores-Sarnat et al., 2003). We recently identified a child with isolated hemimegalencephaly (HMG) caused by a somatic missense (p.E17K) point mutation in *AKT3* present

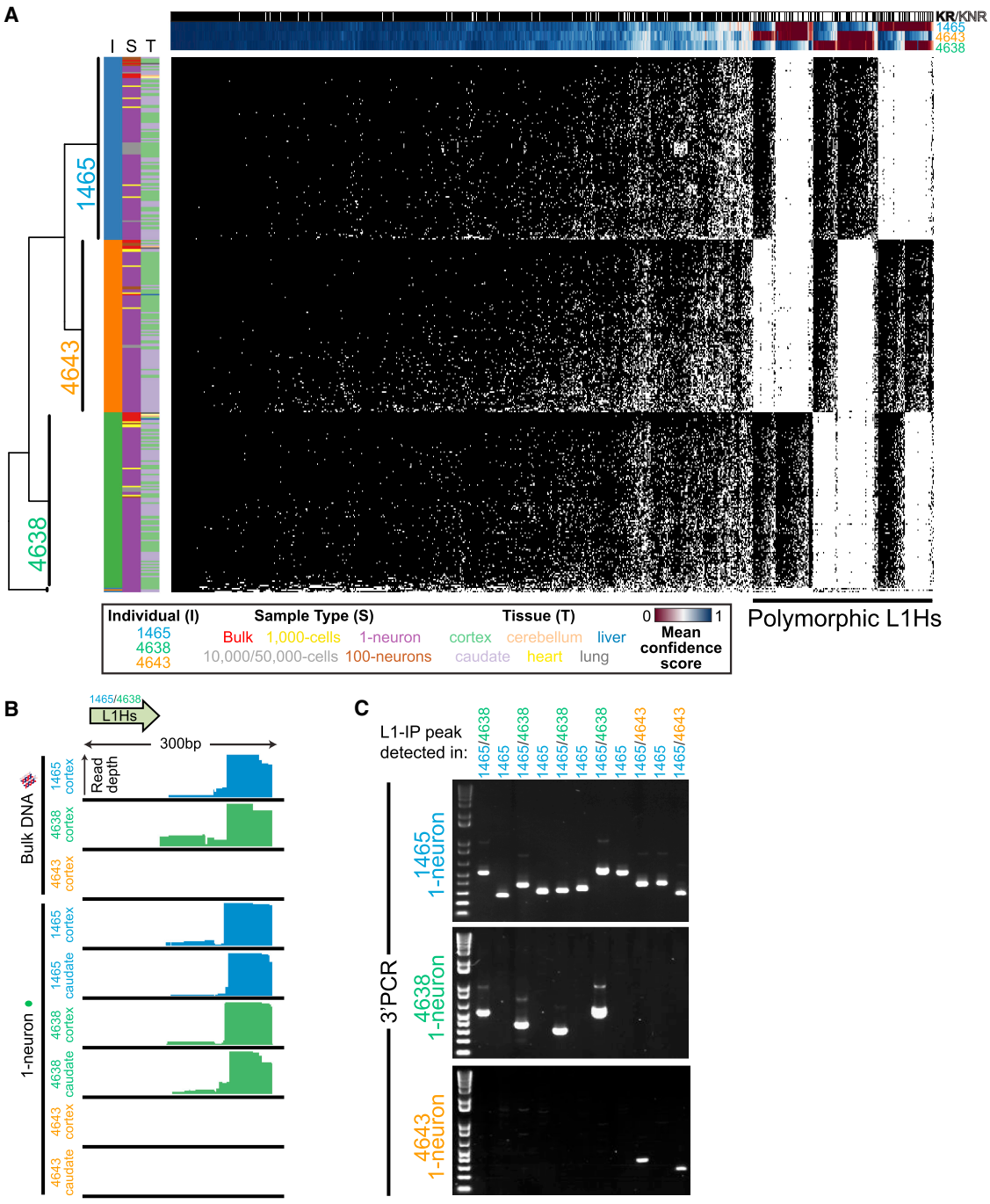

Figure 5. Single-Neuron Fingerprinting with L1-IP

(A) Unbiased hierarchical clustering of all samples sequenced in this study (excluding technical replicates) by transposon profile. Each row represents a sample, and each column represents a specific L1Hs insertion. Data are shown for all KR and KNR insertions with an average score of at least 0.5 in at least one individual's samples. Black and white squares indicate presence or absence, respectively, of the insertion using a confidence score threshold of 0.5. All samples cluster correctly by individual except for three low-quality 1-neuron samples that cluster in a separate branch (bottom branch). Additional row annotations are colored for individual (I), sample type (S), and tissue (T), illustrating correct clustering by individual. Column annotations show annotation for KR (black) and KNR (white) insertions and mean confidence scores across all samples of each individual. Samples also cluster by individual when including all insertions including unknown peaks (data not shown).

(B) L1-IP read coverage for a representative polymorphic known nonreference insertion (L1Hs-KNR-1158).

(C) Representative gel images of 3′PCR of 11 polymorphic germline insertions with 1-neuron DNA. 3′PCR products are only detected in individuals predicted by L1-IP to have the insertion. All polymorphic insertions tested are listed in Table S3.

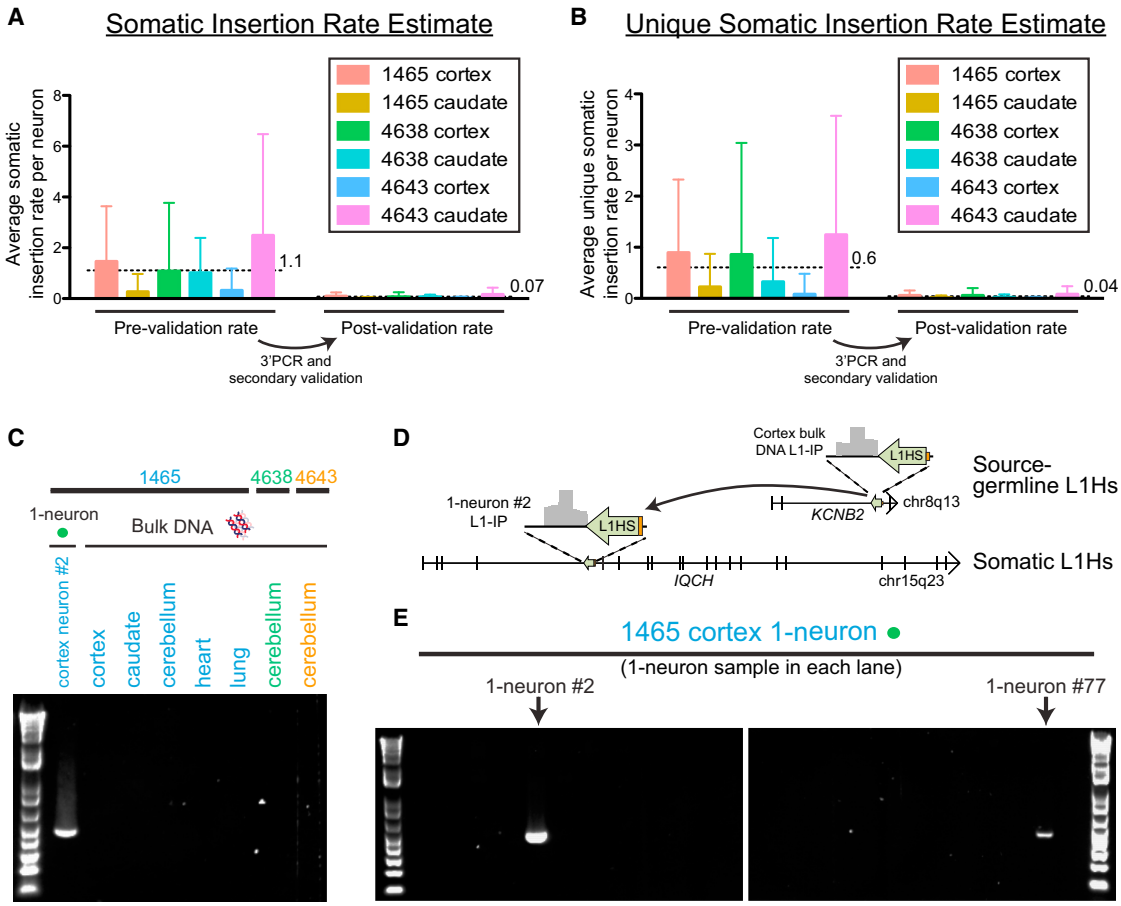

Figure 6. Quantification of Somatic L1Hs Insertions and Validation of a Somatic Insertion in Single Neurons

(A) Mean number (±SD) of somatic insertion candidates per single neuron in each tissue in the study, corrected for sensitivity. The estimated insertion rates per neuron are shown before and after 3′PCR and secondary validation. Horizontal dashed lines and adjacent numbers indicate the mean number of insertion candidates across all single neurons from all tissues. Low-quality samples that did not achieve the necessary KNR detection rate with a confidence score >0.5 were excluded from the analysis in a quality control check ("QC-fail" in Table S2). The number of cells included in each analysis were n = 50, 45, 45, 50, 50, and 44 for 1465 cortex, 1465 caudate, 4638 cortex, 4638 caudate, 4643 cortex, and 4643 caudate, respectively, after removing low-quality samples failing quality control.

(B) Mean number (±SD) of unique somatic insertion candidates (i.e., present in only one single-neuron sample of the individual) per single neuron in each tissue, corrected for sensitivity.

(C) Gel images of 3′PCR validation of a somatic L1Hs insertion found by L1-IP in individual 1465 cortex 1-neuron #2 (L1-IP peak ID chr15_67625710_plus_0_0).

(D) Location of the somatic L1Hs insertion (L1-IP peak ID chr15_67625710_plus_0_0) in antisense orientation in intron 4 of the gene *IQCH* and the corresponding L1-IP peak in 1465 cortex 1-neuron #2. The insertion's target site duplication coordinates are chr15: 67,625,702–67,625,714 (hg19). A 5′ transduction (orange) identified the source L1Hs on chr8: 73,787,792–73,793,823.

(E) Representative gel images from a 3′PCR screen of 83 1-neuron samples from individual 1465 cortex (24 1-neuron samples shown) for the somatic insertion in Figures 6C and 6D. The two cortical 1-neuron samples (#2 and #77) found to have the insertion are shown. 1-neuron #77 was found to have the insertion only in the 3′PCR screen because it was not profiled by L1-IP. 3′PCR product sequencing and full-length cloning confirmed that the insertion had identical 5′ and 3′ breakpoints and TSD in both neurons (#2 and #77).

See also Figure S7 and Table S3.

in the brain, but not in the blood (case HMG-3, Poduri et al., 2012) (Figure 7A). Due to intractable epilepsy, the malformed hemisphere was surgically removed, allowing application of our single-cell method to genotype single sorted cells from this surgical sample and to study the origin of the pathologic cells.

Previous analysis of resected bulk tissue indicated that the mutation was present at ~35% mosaicism based on cloning of PCR products (Poduri et al., 2012). Interestingly, 39% ± 7% (SE; corrected for AD) of single sorted neuronal ($NeuN^+$) nuclei contained the mutation (Figures 7B and 7C and Table S4), similar to the mosaicism in unsorted bulk tissue containing both neuronal and nonneuronal cells. This suggested that the mutation was also present in nonneuronal cells, consistent with the abnormality of both gray matter and white matter in this patient by MRI (Poduri et al., 2012; Figure 7A). Indeed, we confirmed the presence of the mutation in single nonneuronal ($NeuN^-$)

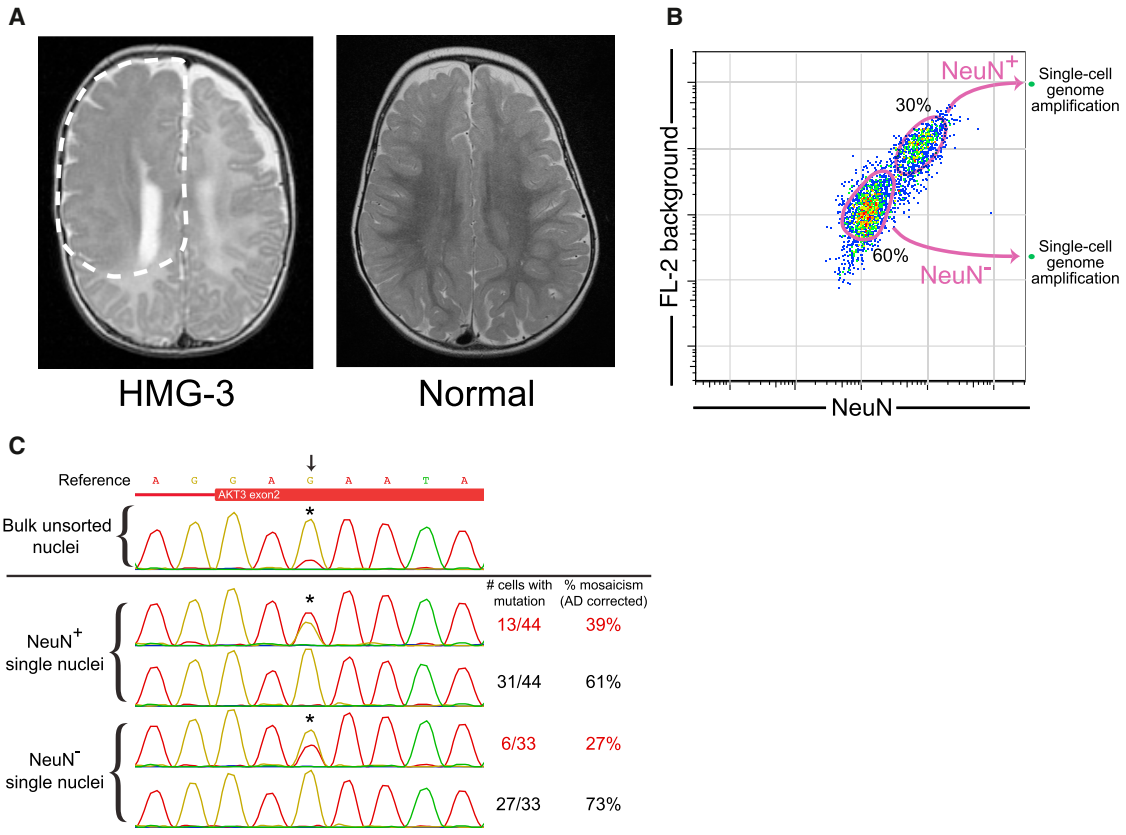

Figure 7. Single-Cell Analysis of a Somatic Brain *AKT3* Mutation Causing Hemimegalencephaly

(A) An axial T2-weighted image from the MRI of the hemimegalencephaly patient, HMG-3, with a somatic AKT3 E17K mutation shows the enlarged right hemisphere with abnormally thick and malformed cerebral gray matter and abnormal signal of the white matter (white dashed line). On the right is an MRI image of a normal brain.

(B) Single-cell FACS sorting of HMG-3 resected cortex.

(C) Representative Sanger sequencing traces of a bulk unsorted nuclei sample and single-cell samples from NeuN$^+$ and NeuN$^-$ populations. The calculated percent mosaicism for single-cell samples (corrected for allelic dropout) is shown. Arrow and asterisks mark the site of the *AKT3* c.49G→A (p. E17K) mutation. See Table S3 for primer sequences, and see Table S4 for percent mosaicism of all samples from HMG-3.

nuclei, at an average percent mosaicism (corrected for AD) of 27% ± 8% (Figure 7C and Table S4). These data indicate that the mutation was present in an early neocortical progenitor capable of giving rise to both neuronal and nonneuronal cells throughout the majority of the hemisphere. The low mosaicism in neurons also indicates that mutant and nonmutant neurons are extensively intermingled in the abnormal hemisphere, presumably reflecting diverse clonal origins of cortical neurons in this pathological condition.

DISCUSSION

Here, we present a single-cell sequencing study of the central nervous system and perform genome-wide analysis to trace patterns of somatic mutation in human brain. We confirmed that somatic retrotransposon insertions can be detected in normal human brain. However, our analysis of L1 insertions found that somatic insertions are rare in normal human cortical and caudate neurons, suggesting that L1 retrotransposition is not a major source of neuronal diversity in cerebral cortex and caudate nucleus. Finally, we used single-cell analysis to study the mosaicism of a somatic *AKT3* mutation, highlighting the potential of single-cell sequencing for cell lineage analysis in human brain.

L1Hs Retrotransposition in Human Cerebral Cortex and Caudate Nucleus

Our validation of a somatic L1Hs insertion with all of the hallmarks of a bona fide retrotransposition event, including a 5′ transduction identifying its source, confirms that somatic L1Hs insertions are present in the normal human brain. The very low-level mosaicism of this insertion and its detection only in cortical neurons further suggest that it may have occurred during cortical development. The source L1Hs on chromosome 8 from which the somatic insertion originated lies in antisense orientation within an intron of the gene *KCNB2* and is a full-length insertion with both open reading frames intact. Although it is present in the human genome reference, it is polymorphic in the population and was present only in individual 1465, but not in the other individuals in this study (data not shown). In addition to this

source L1Hs, only one other L1Hs element has been previously confirmed to be active somatically in humans (Van den Hurk et al., 2007). Further single-cell studies will help to delineate the spectrum of somatic activity of L1Hs elements in different tissues and developmental stages.

Our quantitative analysis of retrotransposition indicates that somatic L1Hs events are rare in adult human cortical pyramidal neurons and caudate neurons. We find that, although we can detect hundreds of known germline insertions in single neurons, >80% of neurons show no unique somatic insertions (i.e., present in one neuron, but not multiple neurons). Somatic L1Hs insertions present in multiple neurons, but not all neurons, as seen for the full-length somatic insertion that we identified, are also rare. On the other hand, we cannot exclude greater rates of L1Hs activity in other cell types or regions of the human brain, or activity of Alu and SVA retrotransposons in the cortex and caudate. Variability in the number of highly active "hot" L1s per individual (Beck et al., 2010) may also lead to variability in somatic retrotransposition rates among individuals; however, the low number of somatic insertions in 300 neurons from three individuals precludes it from being an essential source of neuronal diversity in cortex and caudate that is common in humans.

Our results are generally consistent with the rates of ~1/10,000 to ~1/100 insertion events per human neural progenitor measured in an in vitro $L1_{RP}$ reporter assay (Coufal et al., 2009). This rate is far lower than the rate measured by quantitative PCR (Coufal et al., 2009; Muotri et al., 2010), which estimated a relative copy number increase of L1 of ~5%–10% and an absolute estimate of ~80 somatic L1 insertions per cell in human brain. Studies employing targeted capture of L1 sequences from human brain (Baillie et al., 2011) also reported widespread L1 retrotransposition. These methods are less direct and do not analyze individual neurons but instead analyze pooled DNA from bulk tissue. Compared to sequencing of bulk tissue (Baillie et al., 2011), our approach of single-cell sequencing has the additional advantage that potential artifacts, such as chimeric reads, are easier to recognize because they are present at lower read depth relative to true insertions. The identification of mammalian species that appear to have lost all L1 activity (Cantrell et al., 2008) further suggests that L1 retrotransposition is not a universal requirement for mammalian neurogenesis. Recent L1 profiling of 26 glial brain tumors did not reveal any somatic insertions (Iskow et al., 2010; Lee et al., 2012a), indicating that somatic L1 insertions may be uncommon in glial progenitors as well. Though our study suggests that somatic L1 retrotransposition in the human cortex and caudate is rare, it remains possible that neuronal L1 retrotransposition may occur at higher rates in other brain regions, such as the hippocampus, and/or may play a role as a mutagen in the human brain in neurological disease.

Somatic Mutations Causing Cortical Malformations Can Occur in Neuroglial Progenitors

Our analysis of a somatic retrotransposon insertion and a somatic *AKT3* mutation, each found in more than one cortical neuron as well as at low levels in bulk DNA, suggests that both occurred in progenitor cells of the brain and that other focal brain malformations of unknown etiology may be similarly caused by progenitor mutations during development. The somatic *AKT3* mutation in hemimegalencephalic brain was found in both neuronal and nonneuronal cells, further indicating that the mutation occurred in a neuroglial progenitor. Moreover, the normal-appearing basal ganglia of this patient by MRI (data not shown) would be consistent with a mutation occurring in a neuroglial progenitor in the developing neocortex, but not involving the ventral telencephalon, though caudate tissue was not available for testing.

Our study suggests potential future applications of somatic mutations as cell lineage markers in postmortem human brain. Although retrotransposon insertions appear too rare for systematic study of cell lineages and the specific *AKT3* mutation assayed here clearly changes the behavior of cells carrying the mutation (Poduri et al., 2012), deeper sequencing of single cells might eventually identify diverse, nonfunctional mutations, including mutations at highly mutable sites like microsatellite repeats (Frumkin et al., 2005; Salipante et al., 2008), which may allow more systematic interrogation of lineage relationships even in human postmortem brain.

EXPERIMENTAL PROCEDURES

Full protocols can be found in the Extended Experimental Procedures.

Tissue Sources

Fresh-frozen postmortem tissues of three normal individuals and a trisomy 18 fetus (UMB1465, UMB4638, UMB4643, and UMB866) were obtained from the NICHD Brain and Tissue Bank at the University of Maryland. Hemimegalencephalic brain tissue from case HMG-3 (Poduri et al., 2012) was obtained following neurosurgical resection of the affected right hemisphere.

Single Neuronal Nuclei Isolation and Genome Amplification

Nuclei were purified by sucrose cushion ultracentrifugation and labeled with NeuN antibody (Millipore, MAB377) for flow cytometry as previously described (Matevossian and Akbarian, 2008; Spalding et al., 2005). Single nuclei were sorted with a FACSAria II cell sorter into 96- or 384-well plates and amplified by MDA (Dean et al., 2002). Low-coverage sequencing libraries were made with the NEXTflex DNA-seq kit (Bioo Scientific).

Genome-wide L1Hs Insertion Profiling

L1Hs insertion profiling (L1-IP) libraries were made by modification of the method of Ewing and Kazazian (2010) for a high-throughput workflow and high-level (up to 32-plex) multiplexing. Libraries were sequenced on HiSeq 2000 sequencers (Illumina). A custom data analysis pipeline was created to call and classify L1-IP peaks.

L1Hs Insertion Validation

3′ junction PCR (3′PCR) was performed with one primer specific to L1Hs (L1Hs-AC-22) and a 5′ peak flank primer (upstream to the L1-IP peak) to verify the presence of the predicted insertion. Full-length (long-range) PCR with 5′ and 3′ peak flank primers was performed to clone the entire length of candidate insertions.

ACCESSION NUMBERS

Sequencing data from this study are deposited in the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA056303.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, six tables, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.09.035.

ACKNOWLEDGMENTS

G.D.E. and X.C. performed all experiments, with assistance from L.B.H, P.C.E., H.S.L, J.J.P., and K.D.A. G.D.E. and E.L. analyzed the L1-IP data with input from P.J.P. G.D.E. and X.C. analyzed all other data. G.D.E., X.C., and C.A.W. conceived and designed the project with input from E.C.G and A.P. G.D.E., X.C., and C.A.W. wrote the manuscript.

We thank Peter V. Kharchenko, Tim W. Yu, Vijay S. Ganesh, and Nathan C. Silberman for helpful discussions; Hal Schneider, Richard Bennett, R. Sean Hill, and Christina Kourkoulis for technical assistance; Robert Johnson from the NICHD Brain and Tissue Bank; the Orchestra research computing support team (Harvard Medical School); and the Hematologic Neoplasia Flow Cytometry Core (Dana-Farber Cancer Institute). Brain image in Figure 1A adapted with permission from http://brainmuseum.org, supported by the US National Science Foundation. G.D.E. is supported, in part, by NIH MSTP grant T32GM007753. X.C. is supported, in part, by NIH NIGMS grant T32GM007726. C.A.W. is supported by the Manton Center for Orphan Disease Research and grants from the NINDS (R01 NS079277 and R01 NS035129). C.A.W. is an Investigator of the Howard Hughes Medical Institute.

Received: May 26, 2012
Revised: August 2, 2012
Accepted: September 19, 2012
Published: October 25, 2012

REFERENCES

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534–537.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159–1170.

Blainey, P.C., and Quake, S.R. (2011). Digital MDA for enumeration of total nucleic acid contamination. Nucleic Acids Res. *39*, e19.

Cantrell, M.A., Scott, L., Brown, C.J., Martinez, A.R., and Wichman, H.A. (2008). Loss of LINE-1 activity in the megabats. Genetics *178*, 393–404.

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.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA *99*, 5261–5266.

Erickson, R.P. (2010). Somatic gene mutation and human disease other than cancer: an update. Mutat. Res. *705*, 96–106.

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.

Flores-Sarnat, L., Sarnat, H.B., Dávila-Gutiérrez, G., and Alvarez, A. (2003). Hemimegalencephaly: part 2. Neuropathology suggests a disorder of cellular lineage. J. Child Neurol. *18*, 776–785.

Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U., and Shapiro, E. (2005). Genomic variability within an organism exposes its cell lineage tree. PLoS Comput. Biol. *1*, e50.

Gittins, R., and Harrison, P.J. (2004). Neuronal density, size and shape in the human anterior cingulate cortex: a comparison of Nissl and NeuN staining. Brain Res. Bull. *63*, 155–160.

Gleeson, J.G., Minnerath, S., Kuzniecky, R.I., Dobyns, W.B., Young, I.D., Ross, M.E., and Walsh, C.A. (2000). Somatic and germline mosaic mutations in the doublecortin gene are associated with variable phenotypes. Am. J. Hum. Genet. *67*, 574–581.

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

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., et al. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell *148*, 873–885.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell *141*, 1253–1261.

Kalisky, T., Blainey, P., and Quake, S.R. (2011). Genomic analysis at the single-cell level. Annu. Rev. Genet. *45*, 431–445.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. *7*, 19.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., III, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., et al; Cancer Genome Atlas Research Network. (2012a). Landscape of somatic retrotransposition in human cancers. Science *337*, 967–971.

Lee, J.H., Huynh, M., Silhavy, J.L., Kim, S., Dixon-Salazar, T., Heiberg, A., Scott, E., Bafna, V., Hill, K.J., Collazo, A., et al. (2012b). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat. Genet. *44*, 941–945.

Magavi, S., Friedmann, D., Banks, G., Stolfi, A., and Lois, C. (2012). Coincident generation of pyramidal neurons and protoplasmic astrocytes in neocortical columns. J. Neurosci. *32*, 4762–4772.

Matevossian, A., and Akbarian, S. (2008). Neuronal nuclei isolation from human postmortem brain tissue. J. Vis. Exp. *20*, e914.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. *52*, 643–645.

Mills, S.E. (2007). Histology for pathologists, Third Edition (Philadelphia: Lippincott Williams & Wilkins).

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

Muotri, A.R., and Gage, F.H. (2006). Generation of neuronal variability and complexity. Nature *441*, 1087–1093.

Muotri, A.R., Chu, V.T., Marchetto, M.C.N., Deng, W., Moran, J.V., and Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903–910.

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

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Ovchinnikov, I., Rubin, A., and Swergold, G.D. (2002). Tracing the LINEs of human evolution. Proc. Natl. Acad. Sci. USA *99*, 10522–10527.

Parent, A., and Carpenter, M.B. (1996). Carpenter's human neuroanatomy (Baltimore, MD: Williams & Wilkins).

Poduri, A., Evrony, G.D., Cai, X., Elhosary, P.C., Beroukhim, R., Lehtinen, M.K., Hills, L.B., Heinzen, E.L., Hill, A., Hill, R.S., et al. (2012). Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron *74*, 41–48.

Rehen, S.K., Yung, Y.C., McCreight, M.P., Kaushal, D., Yang, A.H., Almeida, B.S., Kingsbury, M.A., Cabral, K.M., McConnell, M.J., Anliker, B., et al. (2005). Constitutional aneuploidy in the normal human brain. J. Neurosci. *25*, 2176–2180.

Rivière, J.B., Mirzaa, G.M., O'Roak, B.J., Beddaoui, M., Alcantara, D., Conway, R.L., St-Onge, J., Schwartzentruber, J.A., Gripp, K.W., Nikkel, S.M., et al; Finding of Rare Disease Genes (FORGE) Canada Consortium. (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat. Genet. *44*, 934–940.

Salipante, S.J., Thompson, J.M., and Horwitz, M.S. (2008). Phylogenetic fate mapping: theoretical and experimental studies applied to the development of mouse fibroblasts. Genetics *178*, 967–977.

Singer, T., McConnell, M.J., Marchetto, M.C., Coufal, N.G., and Gage, F.H. (2010). LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci. *33*, 345–354.

Spalding, K.L., Bhardwaj, R.D., Buchholz, B.A., Druid, H., and Frisén, J. (2005). Retrospective birth dating of cells in humans. Cell *122*, 133–143.

Van den Hurk, J.A.J.M., Meij, I.C., Seleme, M.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S., et al. (2007). L1 retrotransposition can occur early in human embryonic development. Hum. Mol. Genet. *16*, 1587–1592.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell *150*, 402–412.

Wolf, H.K., Buslei, R., Schmidt-Kastner, R., Schmidt-Kastner, P.K., Pietsch, T., Wiestler, O.D., and Blümcke, I. (1996). NeuN: a useful neuronal marker for diagnostic histopathology. J. Histochem. Cytochem. *44*, 1167–1171.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886–895.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W., and Church, G.M. (2006). Sequencing genomes from single cells by polymerase cloning. Nat. Biotechnol. *24*, 680–686.

EXTENDED EXPERIMENTAL PROCEDURES

Tissue Sources

Fresh-frozen postmortem tissues of normal individuals and a trisomy 18 fetus were obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland (Baltimore, MD). All tissues were frozen at −80°C with postmortem intervals ≤ 5 hr. Case UMB1465 was a 17 year-old male who died in a motor vehicle accident; case UMB4638 was a 15 year-old female who died in a motor vehicle accident; case UMB4643 was a 42 year-old female who died of cardiovascular disease; and case UMB866 was a 21 week-gestation fetus from an electively terminated pregnancy, with a 47XY,+18 (trisomy 18) karyotype. Hemimegalencephalic brain tissue from a 5 month-old male child with a somatic *AKT3* mutation (HMG-3) (Poduri et al., 2012) was obtained following neurosurgical resection of the affected right hemisphere and fresh-frozen at −80°C until use. HMG-3 was an infant with seizures beginning in the first week of life secondary to right-sided hemimegalencephaly. He underwent anatomic hemispherectomy at 5 months of age, which resulted in a dramatic reduction in seizures and developmental improvement. At 10 years of age, he had cognitive impairment but was able to speak fluently and read, and he had left-sided weakness but was able to walk alone. HMG-3 was enrolled in clinical research in accordance with requirements of the Institutional Review Board of Boston Children's Hospital. All tissue samples were confirmed as deriving from the correct individual with AmpFlSTR Identifiler Plus fingerprinting (Applied Biosystems).

Single Neuronal Nuclei Flow Sorting and Labeling

Nuclei were isolated and labeled for fluorescence-activated cell sorting (FACS) based on Spalding et al. (2005) and Matevossian and Akbarian (2008). All procedures were performed at 4°C unless noted. 100-200mg of tissue was homogenized in a dounce homogenizer in lysis buffer (0.1% Triton X-100, 11% sucrose, 5mM $CaCl_2$, 3mM $MgAc_2$, 0.1mM EDTA, 10mM Tris-pH8, 50mM DTT) and ultracentrifuged on top of a sucrose cushion (62% sucrose, 3mM $MgAc_2$, 10mM Tris-pH8, 50mM DTT) at 13,000 rpm for 2 hr in an SW-28.1 rotor (Beckman Coulter). The pellet was resuspended in PBS+3mM $MgCl_2$ solution and filtered through a 40 μm cell strainer. Nuclei integrity, purity and concentration were assessed by light microscopy on a hemocytometer with light trypan blue staining.

For labeling with NeuN, 1.2 μg each of NeuN antibody (Millipore, MAB377) was pre-incubated with Alexa-488 goat α-mouse and Alexa-647 donkey α-rabbit antibodies (Life technologies) in 400 μl PBS+3% BSA+3mM $MgCl_2$ solution at RT for 10 min. Initial experiments were performed with NeuN/Mef2c double-labeling, and since all $NeuN^+$ nuclei were also $Mef2c^+$ (data not shown) subsequent experiments were performed only with NeuN labeling. For NeuN/Mef2c double labeling experiments, 1.2ug each of NeuN and Mef2c antibody (Abcam, ab64644) were used. Alexa-647 secondary antibody was still used in addition to Alexa-488 in NeuN-only experiments to provide background signal (FL-2) relating to nuclear size. 200-500 × 10^3 nuclei were diluted in 1ml PBS+3mM $MgCl_2$ solution and incubated with the antibody mix at 4°C for at least 30 min. Nuclei in Figure S1B were labeled with Alexa488-conjugated NeuN antibody (Millipore, MAB377X).

Single nuclei were sorted at a maximum flow rate of 3.0 with a FACSAria II cell sorter at the Dana-Farber Hematologic Neoplasia Flow Cytometry core into 96-well (qMDA experiments) or 384-well (sequencing experiments) plates. Prior to sorting each plate, the plate holder of the FACSAria automated cell deposition unit was calibrated to the sort stream. Sorting into 384-well plates always left a gap of one empty well in all directions between single cells. 1,000-nuclei samples were sorted into microtubes. We used PBS+3mM $MgCl_2$ sheath fluid for flow sorting and kept flow sorter sample chambers chilled to 4°C to help preserve nuclear integrity. SSC-H versus SSC-W and FSC-H versus FSC-W doublet discrimination gates, and a stringent '0/32/16 single-cell' sort mask, were used to ensure one and only one nucleus was sorted per well. Initial experiments with DRAQ5 staining confirmed exclusion of doublets. A sorting purity quality control was performed for every sample prior to sorting for amplification. In the sorting purity quality control, a portion of the purified nuclei was sorted, and the sorted nuclei were reanalyzed by flow cytometry to ensure that nuclear integrity was preserved and that sort purity was > 98%. Single neurons were sorted for each tissue into 1 or 2 plates, one of which contained negative (0 nuclei) and positive control (10- and 100-nuclei) wells. Amplified human DNA was never observed in negative control wells in quality control assays. Successful sorting of single nuclei into 384-well plates was monitored by quantitation of yield and with multiplex PCR (see below). Every plate's sorting success rate per well was > 80%, and the average sorting success rate per well across all plates was 94%.

RT-PCR and Western Blots

5,000 nuclei from each population were sorted into microtubes. RNA was extracted using RNeasy (QIAGEN) and cDNA synthesized with the Superscript III First-Strand Synthesis System (Life tech.), and in separate experiments protein lysates were used for Western blots. RT-PCR primers for neuronal (*SNAP25* and *SYT1*) and nonneuronal (*GFAP* and *AQP4*) transcripts (Lyck et al., 2008; Nielsen et al., 1997; Sudhof, 2004), and an input control transcript (*RPL37A*), were designed with Primer3 (Rozen and Skaletsky, 2000) to span introns and are listed in Table S3. Antibodies used for Western blots were NeuN (Millipore, MAB377), Olig2 (Millipore, AB9601), and HistoneH3 (Abcam, ab1791) at 1:1000 dilutions.

Single-Neuron Genome Amplification

All work was carried out in a UV-treated laminar flow cabinet, and all surfaces, plastics and non-biologic buffers were UV-treated at least 30 min. Reagents were added without touching the liquid surface to avoid losing parts of the single genome. Nuclei sorted into

384-well plates were sorted into 2.8 μl lysis and denaturing buffer (200mM KOH, 5mM EDTA, 40mM DTT), and neutralized with 1.4 μl neutralization buffer (400mM HCl, 600mM Tris-pH7.5). 15.8 μl MDA reaction-mix (recipe below) was added to each well and incubated in a thermal cycler at 30°C for 16 hr (no lid heating), followed by 3 min. at 65°C.

MDA reactions (Dean et al., 2002) were optimized for hexamer, dNTP, and phi29 polymerase concentrations by amplifying control human bulk DNA and assaying yield with Quant-iT Picogreen (Life tech.). All reaction conditions were confirmed to have high-molecular weight (>30kb) products by standard and alkaline gel electrophoresis (data not shown). Following optimization of MDA reaction conditions, MDA reagent concentrations used in reactions in this work were as follows: 1x RepliPHI phi29 reaction buffer (Epicentre), 50 μM random hexamer 5′-dNdNdNdN*dN*dN-3′ (* = thiophosphate linkage) (IDT, Inc), 2mM each dNTP, 40U RepliPHI phi29 polymerase (Epicentre), and nuclease-free UV-treated water. Quantitative MDA reactions (Zhang et al., 2006) were monitored on a StepOnePlus real-time PCR instrument (Applied Biosystems) by addition of 0.1x SYBR Green I (Life tech.) and fluorescence was measured every 6 min. for 7 hr.

Following MDA amplification, 0.5 μl of reaction products were diluted 1:50 for Picogreen, multiplex PCR, and Identifiler quality control assays (see below). The remainder of the DNA was purified with AMPure XP beads (Beckman Coulter), treated with 10U mung-bean nuclease (NEB) at 30°C for 30 min to debranch the MDA product structure (Zhang et al., 2006), cleaned-up with the DNeasy Blood & Tissue 96 kit (QIAGEN) skipping the tissue digestion protocol steps, and again assayed for final yield with Picogreen.

Amplified Genome Quality Control

Dilutions (1:50) of MDA reaction products prior to cleanup were used for quality control assays. Every MDA reaction well, including negative and positive controls, was assayed with 2 methods: (a) Picogreen quantitation to measure yield and confirm success of controls (negative control reactions produce about one-half the yield of single nuclei reactions) and determine into which wells a nucleus was successfully sorted; (b) multiplex PCR for 4 arbitrarily chosen loci from different chromosomes in the human genome: to exclude human DNA contamination in negative controls reactions, independently determine which wells contain a successfully sorted nucleus, and exclude failed nuclei amplifying < 3 loci, likely indicating loss of significant parts of the genome during sorting or amplification. 96.8% of the wells into which a cell had been successfully deposited (at least one of the 4 loci amplified) passed our 4-locus multiplex PCR quality control (3 or 4 of the 4 loci amplified). Negative control wells and wells into which a nucleus failed to sort always produced both low yield by Picogreen and none of the multiplex PCR bands. Multiplex primers were designed with PrimerStation (Yamada et al., 2006). Multiplex PCR reactions contained 5 μM of each primer (primers listed in Table S3), 1x HotStarTaq reaction buffer (QIAGEN), supplemental 1.5mM $MgCl_2$, 0.2 μl HotStarTaq polymerase (QIAGEN), 0.4mM dNTP, and 2 μl of 1:50 MDA reaction product, in 20 μl reaction volumes. Thermal cycler conditions were: 94°C 15 min, (94°C 1min, 68°C 1 min decreasing by 1°C every cycle, 72°C 1min, for 13 cycles), (92°C 1min, 55°C 1 min, 72°C 1 min, for 27 cycles), 72°C 10 min.

To further confirm the absence of any human DNA contamination and confirm identity of sorted nuclei, additional quality control on a subset of 8-16 wells, including negative and positive controls, from each sorted plate, was performed by Identifiler muliplex genotyping of 16-microsattelite (STR) loci with the AmpFlSTR Identifiler Plus kit (Applied Biosystems) on a 3130xl Genetic Analyzer (Applied Biosystems). 1:50 dilution plates from above were further diluted to ~0.1ng/ul based on Picogreen quantitation for use in Identifiler assays. Unamplified bulk DNA genotypes were used as a reference. Loci homozygous in an individual were excluded from preferential amplification (PA), low-amplification (LA), and allelic dropout (AD) calculations since they cannot be used to estimate per-allele PA, LA, and AD. 14 heterozygous loci were included in analysis for individuals 1465 and 4643 and 11 loci for 4638 (i.e., 11-14 heterozygous loci assayed in 92 single neurons = 1,183 loci assayed for the 1-neuron group). Genotypes of all samples were checked for concordance to the bulk reference genotype of the individual. Preferential amplification was defined as loci where the area under the trace of one allele was > 3x the area under the trace of the other allele. Low amplification was defined as callable alleles with traces of area < 1,000 fluorescence units. Identifiler fingerprinting was performed on a subset of nuclei from every plate of nuclei sorted in this work.

Affymetrix SNP6 microarray genotyping was performed (by Expression Analysis, Inc.) on bulk DNA from cortex and lung tissue from individuals 1465, 4638 and 4643, and 3 single cortical neurons from individual 1465. Genotypes were called using the Affymetrix Genotyping Console with the Birdseed-v2 algorithm with standard settings. Genotypes called with confidence scores ≤ 0.01 in both the sample and reference being compared were used for analysis. Genotype concordance was calculated as: (# of loci with AA calls in both samples + # of loci with AB calls in both samples + # of loci with BB calls in both samples) / (total # of loci). The fraction allelic dropout (dropout rate) was calculated as: (# of heterozygous AB loci in the reference with AA or BB calls in the sample) / (2 × # of heterozygous AB loci in the reference). The fraction of discordant alleles was calculated as [(2 × # of loci with BB calls in the sample + # of loci with AB calls in the sample, where the reference is AA) + (2 × # of loci with AA calls in the sample + # of loci with AB calls in the sample, where the reference is BB)] / (2 × # of AA + 2 × # of BB loci in the reference). Depending on the number of loci passing the confidence score threshold, for single-neuron versus bulk DNA comparisons between 250,000 and 350,000 loci were included in each comparison for genotype concordance, 60,000–75,000 loci for allelic dropout, and 200,000–300,000 loci for discordant allele calculations.

Whole-Genome Sequencing Libraries

Whole-genome sequencing barcoded libraries for low-coverage sequencing were prepared from 1 μg of DNA with the NEXTflex DNA sequencing kit (Bioo Scientific), for multiplexed sequencing at the Harvard Biopolymers Facility (Harvard Medical School) on a HiSeq

2000 sequencer (Illumina). Library details are in Table S2. Sequencing data is available at NCBI SRA with accession number SRA056303.

Sequencing Copy Number Analysis

Raw reads were trimmed at the 3′ end of low quality-score sequence and mapped to hg18 with Bowtie (Langmead et al., 2009) with -v 2 -m 1--best --strata settings. Chromosome copy numbers (Figure 2B) for each chromosome were calculated as the fraction of reads in each sample aligning to the chromosome, normalized to the median fraction of reads aligning to the chromosome across all 8 neurons. For autosomal chromosomes, these median-normalized relative chromosome copy numbers were multiplied by 2 to obtain absolute chromosome copy numbers, since both individuals were confirmed to have 46XY and 47XY,+18 karyotypes (i.e., 2 copies of each autosome except for chr18 in the trisomy 18 individual). The normal 46XY karyotype of the normal individual (UMB1465) was confirmed by comparison of 1465-bulk DNA low-coverage sequencing samples to in silico simulated reads from a 46XY genome (data not shown). The trisomy 18 individual (UMB866) was confirmed to have a 47XY,+18 karyotype from clinical karyotyping data at the NICHD Brain and Tissue Bank.

Higher-resolution copy number analysis was performed by creating 6,000 (~500kb) bins spanning the entire genome with boundaries defined so that each bin contains an equal number of reads, B_{Ref}, in the reference sample (Navin et al., 2011). The number of reads in the sample being analyzed in each bin, n, defined by the reference were then counted ($B_{Sample,n}$) and normalized to the sample's total read depth, T_{Sample}. B_{Ref} was also normalized to the reference sample's total read depth, T_{Ref}. For any bin 1…n, the relative copy number was then calculated as $CN_n = (B_{Sample,n}/T_{Sample})/(B_{Ref}/T_{Ref})$. Technical replicate low-coverage whole-genome sequencing libraries from the 100-neuron #1 (4 replicates) and 100-neuron #2 samples (3 replicates) were highly reproducible ($R^2 > 0.9$ for all pair-wise comparisons of copy-number profiles) and therefore pooled for analysis. Samples being analyzed were always excluded from the reference samples used for normalization. Bin copy numbers were further normalized to the global median of all bins with copy number > 0.5 since dropout leads to an upward shift of the global midline (dropout bins lead to a proportional increase in reads in normal copy number bins in a given sample; global median normalization corrects this by shifting down all data points by an equal amount). The final relative copy numbers were $\log_2$ transformed for analysis in R (R Development Core Team, 2011) and visualization with scripts modified from the aCGH package.

Genome-wide L1 Insertion Profiling Libraries

L1 insertion profiling (L1-IP) libraries were made by modification of the protocol of Ewing and Kazazian (2010) for a high-throughput workflow and high-level (up to 32-plex) multiplexing. The principle of the L1-IP method developed by Ewing and Kazazian (2010) is two nested PCR reactions each with a primer targeting nucleotides in the L1 3′UTR specific to L1Hs (L1Hs-AC and ILMN-Adaptor1_L1Hs-G targeting positions 5930-5931 and 6015, respectively in the LRE-1 sequence) (Dombroski et al., 1991; Ovchinnikov et al., 2002). L1Hs-AC is paired with 8 arbitrary 5bp seeds (seed primers) in the first PCR, and an oligo (ILMN-SeqAdaptor2) incorporating an Illumina sequencing adaptor is used in the second PCR. Modifications to the protocol published by Ewing and Kazazian (2010) were as follows: (a) Barcoding for multiplex sequencing: Library seed primers were modified to include 5-base (pilot experiments) or 6-base (all subsequent experiments) barcodes preceded and followed by 5 degenerate N nucleotides to avoid barcode interference with sequencer cluster position calling with lower level multiplexing and to avoid barcode biasing of primer annealing during library PCR. Primers used in library preparation are listed in Table S3. Bulk DNA libraries and pilot libraries (1465-cortex unsorted 50,000-nuclei, 100-neuron, and neurons #1-8 samples) were sequenced at 4- or 8-plex, and remaining libraries at 16- or 32-plex. (b) Cleanup of round 1 PCR reactions was performed with AMPure XP beads (Beckman Coulter). (c) Gel size selection: SYBR Safe DNA stain (Life tech.) and a blue light transilluminator were used for gel electrophoresis visualization to avoid operator exposure to UV during gel size selection. Gel bands were cut with clean scalpels and spatulas for each 8 seed primer PCR products belonging to each sample. (d) Gel extraction: DNA was extracted from gel bands in a high-throughput format with the QIAEX II Gel Extraction kit (QIAGEN) in 96-well deep-well plates. (e) Seed reaction pooling: The 8 different seed reaction products for each sample were quantified by qPCR using Illumina library adaptor primers and pooled in equimolar amounts. (f) Sample library multiplexing: after seed reaction pooling, each sample was quantified again by qPCR with Illumina library adaptor primers and samples prepared for sequencing in the same lane were multiplexed in equimolar amounts, followed by Pfu end-repair and MinElute (QIAGEN) cleanup. Final library size distributions were assayed on High Sensitivity DNA bioanalyzer chips (Agilent).

All work was performed in 96-well plates with 8 samples × 8 seed primers each = 64 reactions per plate. Samples with identical barcodes were processed in separate plates and gels at all steps of the library preparation. 100ng of DNA was used for each of the 8 seed reactions, such that 800ng of DNA was required per sample. Libraries were sequenced at the Harvard Biopolymers Facility (Harvard Medical School) and the Tufts University Genomics core facility on HiSeq 2000 sequencers (Illumina). Library details are in Table S2. Sequencing data is available at NCBI SRA with accession number SRA056303.

L1 Insertion Profiling Library Analysis

A schematic of the pipeline for analysis of L1 Insertion Profiling libraries is shown in Figure S3D. The pipeline was created with custom bash shell, Perl and R (R Development Core Team, 2011) scripts, and was run on the Orchestra research computing cluster (Harvard Medical School) and local computers. Parts of the pipeline made use of BEDTools (Quinlan and Hall, 2010) and SAMtools (Li et al.,

2009). Read coverage visualizations (Figures 3B, 5B, 6D, and S7E) were made with the Integrative Genomics Viewer (Robinson et al., 2011). Figure 4 was made with Circos (Krzywinski et al., 2009).

Read Alignment

Raw reads from each lane were assessed for per-base and other quality characteristics with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), and the initial 5 bases (degenerate 'N' bases in the library adaptor) were trimmed. Individual sample libraries were then demultiplexed according to the barcode sequence using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit), allowing for 1 mismatch. The barcode and the subsequent 10 bases (second set of 5 degenerate 'N' bases in the library adaptor and 5bp seed sequences which may contain mismatches), were trimmed. Poor quality sequence at the 3′ end of the reads (between 10 to 17bp), based on the FastQC quality control, were also trimmed. Reads were then aligned to hg19 with Bowtie (Langmead et al., 2009) with the -v 1 -m 1 settings which allows one mismatch and at most one alignment to the genome (multi-mapping reads are discarded). These settings were optimized to minimize false positive alignments. Across all samples, 46% ± 6% of reads aligned on average to the genome in this initial mapping (43% ± 4%, 49% ± 7%, and 45% ± 6% mean ± SD, respectively, for bulk DNA, non-single-cell amplified DNA samples, and single-cell samples). Reads that did not align in the initial mapping with 6 or more trailing 'T's were recovered in order to capture reads that extend into the L1Hs poly-A tail. The trailing 'T's were trimmed and reads at least 20bp long were realigned to the genome with Bowtie with the -v 1 -m 1 settings. On average, 21% of recovered polyT reads were successfully mapped after trimming.

Low-levels of cross-contamination between samples was observed, i.e., some insertions present in the germline of only one individual were also found at very low levels in samples from different individuals multiplexed on the same sequencing lane. This cross-contamination was confirmed to have occurred on the Illumina sequencer itself (i.e., between samples prepared separately and pooled only prior to sequencing). This has been previously observed by others (Kircher et al., 2012) and is likely due to low-quality 'mixed clusters' on the sequencer containing two different templates. This unavoidable and previously unknown artifact of Illumina sequencers would affect any Illumina multiplex sequencing experiment. However, it did not affect our data analysis as this was only found for a small number of insertions and always at significantly lower read depth in the affected samples relative to samples that have the insertion.

Peak Calling

Peak calling is performed separately for reads aligning to the plus and minus strands of the reference genome, since the sequencing libraries preserve information regarding the insertion direction. Peak calling is further performed separately for reads mapped in the initial alignment (main peaks) and for polyT-trimmed reads (polyT peaks). Peak calling is performed by searching for positions in the genome with $\geq$ 5 reads (for samples multiplexed at 32-plex per lane), or $\geq$ 10 reads (for samples multiplexed at < 32-plex per lane). When a position with the threshold number of reads is found (i.e., $\geq$ 5 or $\geq$ 10), a peak call is triggered. The peak is extended as long as the read coverage is above the threshold, while allowing up to 500bp gaps below the threshold for main peaks, and up to 100bp gaps for polyT peaks. The total number of reads aligning to the peak is then counted and normalized to the total sample mapped read depth (reads per million mapped reads, RPM). The number of unique start positions of reads aligning in the same direction as the peak are also counted. PolyT peaks are then joined and annotated to the main peaks if their peak position plus 100bp upstream from the start of the peak overlaps a main peak. Final peak calls have a total of 4 features: main peak read depth, main peak number of unique read start positions, polyT peak read depth, and polyT peak number of unique read start positions.

Peak Classification and Somatic Insertion Analysis

Somatic insertion analysis took place in four steps: (a) Initial joining of each sample's peaks to the bulk DNA samples of the individual, and further peak annotation (see 'Peak annotation' below). (b) Peak classification that scores peaks based on the number of reads and number of unique read start positions. (c) Selection of peaks from samples for somatic insertion calling based on confidence scores or other criteria. (d) Somatic insertion calling.

Details of each of these steps follows. Note that steps (a) and (d) require joining of peaks across samples, since coordinates of peaks from the same insertion are not precisely identical in different samples. Joining peaks across samples allows comparison of peaks corresponding to the same insertion, but which do not have precisely identical peak coordinates.

Step (a)

Each sample was joined separately to the bulk DNA samples from the same individual. For example, peaks from UMB1465 single cortex neuron #1 were joined to peaks from all UMB1465 bulk DNA samples. Joining each non-bulk DNA sample separately to the bulk DNA samples, rather than joining all non-bulk DNA samples together, avoids the potential for joining unrelated peaks with large numbers of samples. The joining process also performs annotation of known insertions and filters non-L1Hs L1Pa insertions. See below 'Peak joining across samples' for details.

Step (b)

As with any genomic sequencing method, a classification model is necessary to differentiate signal from noise. Therefore, in the second step we implemented two peak classifiers, one that is run on the bulk DNA samples, and one that is run on the non-bulk DNA sample that was joined to the bulk DNA samples: (1) A *bulk DNA classifier* is first run on the individual's bulk DNA samples. This classifier generates true/false labels for each peak in the individual for use as a reference in the second classifier of the non-bulk DNA (e.g., 1-neuron) sample. (2) A *non-bulk DNA classifier* is run on the non-bulk DNA sample, using the labels generated in the bulk DNA classifier.

Both classifiers produce confidence scores between 0 and 1, which rank the peaks from least to most likely to be a true L1Hs insertion.

Bulk DNA Classifier. Peaks that are reproducibly detected in both bulk DNA L1-IP library technical replicates from a given tissue of an individual are predicted to be more likely to be true insertions. The bulk DNA classifier performs iterative regressions that optimize the reproducibility of detection between technical replicates, and is run separately on each pair of tissue bulk DNA technical replicates of the individual. The classifier uses the $\log_2$-transformed normalized read depth and the number of unique read start positions per peak as features. Bulk DNA tissue technical replicate pairs were available for all bulk DNA samples except for 4638-lung.

In an initialization step, a threshold of normalized read depth (RPM) of 3 (i.e., $\log_2$-transformed read depth of 1.58), and number of unique read start positions of 2 is used to initialize labels as true or false using peak data from replicate #1. A logistic regression model is then created using peak data from replicate #2 modeled on these initial labels. This model is then applied to the data from replicate #1, and peaks with a score (logistic regression output) of > 0.5 are labeled as true. This process is iterated, wherein in each iteration a logistic regression model is built using the data from replicate #2 and the updated labels, followed by application of the model to the data from replicate #1 to update peak labels using a score threshold of 0.5. This process is iterated until the improvement of the likelihood of the logistic regression model is < 5% (i.e., convergence). Separately, this iterative modeling is performed in reverse, using labels initialized from data from replicate #2, then building models using data from replicate #1, and applying the models to data from replicate #2. After both iterative modelings converge, the final *reproducibility score* for each peak is calculated as the average score of the peak from both of the converged models.

On average across all bulk DNA samples, 96.9% ± 0.6%(SD) of peaks with reproducibility score > 0.5 were KR or KNR peaks, indicating that the bulk DNA classifier has a high specificity in discriminating true L1Hs insertions. For individuals 1465, 4638, and 4643, respectively, 655 ± 13(SD), 661 ± 18, and 655 ± 24 KR insertions, and 102 ± 5, 101 ± 8, 103 ± 8 KNR insertions were detected with a score > 0.5. This indicates that the bulk DNA classifier also achieves a high sensitivity as these are approximately the expected numbers of KR and KNR L1Hs insertions present on average in an individual (Ewing and Kazazian, 2010). After running the bulk DNA classifier on all the bulk DNA replicate pairs of the individual, the set of peaks in which at least one bulk DNA replicate pair had a reproducibility score > 0.5 were labeled as 'true' for the subsequent non-bulk DNA sample classifier.

Non-bulk DNA Classifier. The second classifier was run on the non-bulk DNA sample (e.g., 1-neuron sample) that was joined to the bulk DNA samples. A logistic regression modeling is performed using peak feature data (normalized $\log_2$-transformed read depth and unique read starts) of the non-bulk DNA sample versus the peak labels from the above *bulk DNA classifier*. The confidence score is the score output by the model. All peak features and classifier results were then loaded into a database.

Step (c)

Subsets of peaks from samples necessary for the somatic insertion analysis (e.g., peaks from 1-neuron and bulk DNA samples) could be selected based on classifier score cutoffs and any other parameters and annotations.

Step (d)

In the fourth step, selected peaks were joined in the same way as in step (a) and reannotated to known L1Hs peaks, forming an 'analysis batch'. L1Pa filtering (see below) was not performed in this step since it was already performed in step (a). Within an analysis batch, samples were categorized into multiple potentially overlapping 'groups' corresponding to, for example, the tissue (e.g., cortex, heart, etc.), the sample type (i.e., bulk DNA, 100-neuron, 1-neuron), and the cell type (i.e., neuron or glia), and combinations of these (i.e., cortex 100-neuron samples). Per-group statistics were calculated for each peak (e.g., fraction of samples in the group with the peak, fraction of samples in the group with the peak above a score cutoff, mean read depth, mean unique read starts, etc.), which aided in later data analysis. Somatic insertion candidates are peaks detected in a single-neuron sample, but not the germline reference, using a confidence score threshold that achieves 50% sensitivity for KNR reference insertions and excluding peaks within 20kb of known (KR/KNR) insertions. The germline reference was defined as insertions present in at least one bulk DNA or unsorted 50,000-nuclei sample (confidence score thresholds 0.1 and 0.3, respectively) from all tissues of the individual except the tissue from which the single neuron derived. Low-quality outlier samples that did not achieve the necessary KNR detection rate with a confidence score > 0.5 were excluded from the analysis in a quality control check ('QC-fail' in Table S2).

Peak Joining across Samples

To join peaks across samples, all peaks in samples to be joined that are within 100bp of each other and on the same strand are merged to create a merged peaks reference. The merged peaks reference is annotated for known L1Hs insertions, and older (non-L1Hs) L1Pa insertions are filtered (see 'Peak annotation' section). The merged peaks reference is then individually intersected with each sample's peaks in a strand-specific manner, and each peak in the merged peaks reference is annotated with the ID and feature data of peaks in the samples that overlap it. In rare situations when more than one sample peak overlaps a peak in the merged peaks reference, the peak with the higher read depth is retained. Hierarchical clustering by L1 insertion profile (Figure 5A) was performed with the heatmap.2 function in R.

Peak Annotation

Individual sample peaks and peaks in the merged peaks reference are annotated with additional features to facilitate analysis. Sample peaks are annotated with the distance to the nearest peak in that sample and its ID, and the distance to the nearest KR or KNR insertion detected in at least one bulk DNA sample. Peaks in the merged peaks reference are further annotated as follows:

Known L1Hs Annotation. Annotation of peaks as known KR or KNR insertions is made if the peak is in opposite orientation to the known insertion, and a region from the peak's start position to 600bp downstream from the center of the peak (or to the 3′ end of the peak if this is farther downstream) overlaps ± 10bp of the last 3′ base of the known insertion. The KR and KNR insertions used for annotation are in Table S5. Peaks were annotated as KR according to hg19 L1Hs insertions in RepeatMasker (Smit et al., 2010; version 3.3.0, Library version 20120124). The 'out' file of RepeatMasker was pre-processed and further annotated in the following way before use: 'split' L1Hs insertions are reference insertions split into two by intervening sequence (e.g., Alu insertion or genomic deletions). 'Split' insertions are marked by RepeatMasker and were therefore collapsed into one insertion based on the RepeatMasker insertion ID. The collapsed insertions were annotated with the strand of the part containing sequences matching the L1-IP library primer sites since this is the directionality that would be detected by the L1-IP method. The collapsed insertions were also annotated with the L1Hs consensus alignment coordinates spanning both parts. Joining the boundaries of the two parts of split KR insertions took into account special cases, for example: split inversion insertions where each of the two parts faces a different direction and each has one or both L1-IP library primer sequences (theoretically leading to two L1-IP peaks for one insertion); merging of an insertion split into 3 parts; and split insertions skipping the L1-IP primer sites where joining the boundaries of the two parts would have falsely indicated the presence of L1-IP primer binding sites. Each KR insertion was then categorized into one of 4 categories (see Table S5 for category annotation) specifying whether it aligned to the L1Hs consensus sequences corresponding to the last 15bp of the L1-IP library primers (positions 6037 to 6051 for L1Hs-AC and 6121 to 6135 for ILMN-Adaptor1_L1Hs-G primer in RepeatMasker L1Hs consensus coordinates). Category 1 indicates alignment to neither, category 2 only to position 6037-6051, category 3 only to position 6121-6135, and category 4 to both. Only KR L1Hs insertions aligning to both of these regions of the L1Hs consensus would be detectable by the method, since older 3′ truncated L1Hs insertions (categories 1, 2 and 3) would not have the sequences necessary for both the L1-IP library primers (Figure S3A). KR insertion sequences were also aligned to the L1Hs consensus sequence in ClustalW2 (Larkin et al., 2007) and annotated as to whether they contained an exact match to the AC and G L1Hs diagnostic nucleotides that are targeted by L1-IP (see Table S5). Peaks in samples were annotated as KNR insertions if they corresponded to an insertion found in one of several previously published large-scale population studies of nonreference L1 insertions, specifically: dbRIP L1Hs insertions (Wang et al., 2006); novel insertions found in more than one normal sample by Iskow et al. (2010) (Table S2); Table S1 from Ewing and Kazazian (2010); insertions found in at least 5 individuals in table S4 of Ewing and Kazazian (2011); table S1 sequence verified insertions in Huang et al. (2010); and L1 insertions validated as part of the 1000 genomes project in table S1 of Stewart et al. (2011). KR insertions were filtered from this list, coordinates converted to hg19 as necessary, and strand information for insertions in Iskow et al. (2010) were obtained as the opposite of the strand found by Blat of the genomic flanking sequence. Insertions called by more than one study within 250bp and on the same strand were merged. Insertions called with different strands in different studies were merged only if one strand was annotated in more studies than the other strand.

L1Pa Filtering. Despite the high specificity for L1Hs of the L1-IP method of Ewing and Kazazian (2010) adapted here, libraries still amplify some older primate-specific inactive L1s (L1Pa) at low levels (average 499 L1Pa peaks per 1-neuron sample). L1Pa and L1P1 peaks were retrieved from RepeatMasker (L1PA* and L1P1) (Smit et al., 2010) (version 3.3.0, Library version 20120124) to assist in filtering these. While Ewing and Kazazian (2010) filtered against the set of all L1Pa annotations, our filtering was much less stringent in order to avoid falsely filtering true L1Hs peaks. Rather than filtering against all 125,614 L1Pa genomic positions, only 5,117 L1Pa's detected in at least two samples in a set of 46 bulk DNA samples and 8 unsorted 50,000-nuclei MDA samples (with a threshold for peak calling of 5 reads) were used for filtering. Peaks were annotated as an L1Pa peak if they were in opposite orientation to an L1Pa in the L1Pa filtering set, and a region from the peak's start position to 600bp downstream from the center of the peak (or to the 3′ end of the peak if this is farther downstream) overlaps ± 10bp of the last 3′ base of the L1Pa. Rare instances of peaks corresponding both to hg19 reference L1Hs and L1Pa insertions annotated in Repeatmasker were not filtered as these are more likely to derive from the reference L1Hs insertion than the adjacent L1Pa insertion.

An additional analysis of seed primer sequence performance was performed:

Seed Sequences. All 5bp sequences found in the seed region of all reads in 16 bulk DNA, 11 unsorted 50,000- nuclei, 9 100-neuron and 20 1-neuron samples were counted for analysis of the L1-IP seed primer performance. The set of 8 5bp sequences used in the library were the top 8 seeds found most frequently in the sequence data in all sample types and accounted for 76% of the reads, confirming that the seed primers used in library PCR mostly hybridize to their specific 5bp sequences. An additional 34 overrepresented seeds, most of which are 1 base mismatched from the 8 seeds used in the library seed primers, accounted for 90% of reads captured by the method. The relative abundance of seeds recovered for each of the four sample types was highly correlated (R^2 > 0.97 for all pair-wise comparisons).

L1Hs Insertion Validation

Batch Primer Design

A custom primer design pipeline for L1Hs insertion validation was programmed in Excel, Galaxy (Goecks et al., 2010), and Perl. L1-IP peak coordinates were used to define 750bp flanks 5′ and 3′ of the L1-IP peak, in order to design primers that flank the candidate insertion. The L1-IP peak 5′ flank coordinates used to search for primers were 800 to 50bp upstream of the 5′ end of the peak, and the L1-IP peak 3′ flank coordinates used were 400 to 1150bp downstream of the 3′ end of the peak. For peaks matching KR insertions, the L1-IP peak 3′ flank coordinates were the 750bp upstream of the 5′ end of the KR insertion. These flank coordinates were used to

extract genomic sequences both from an unmasked hg19 reference and an hg19 reference masked for non-unique 20bp sequences using the Duke Uniqueness track available in the UCSC genome browser. Both sets of hg19 references were used in parallel to search for high quality PCR primers in each flank using Primer3 (Rozen and Skaletsky, 2000), with the following settings: target product size range: 301-850bp with preference for shorter amplicons, minimum primer size of 20bp, and a check for human repeat mispriming (remaining parameters are the default settings in the web version of Primer3). The batch primer design scripts then perform additional quality control on Primer3 primer results by checking the number of hits of each primer in the genome and the number of predicted PCR products using the Blat and in silico PCR functions of the UCSC genome browser (Kent et al., 2002). Primers with 1 genome hit and 1 predicted product were chosen from the unmasked reference primer design results. Primers for peaks without primer pairs matching these criteria were chosen from the uniqueness-masked reference primer design results, again allowing at most 1 genomic hit for the primers and 1 predicted product. Primers for peaks still remaining without an adequate primer pair were then chosen allowing the L1-IP peak 3′ flank primer to have 2 genomic hits. Primers for peaks still remaining without adequate primer pairs were manually designed with the aid of the Duke uniqueness track and Primer3. All primers were purchased from IDT.

3′ Junction PCR and Full-Length PCR Validations

Two types of PCR were used for L1Hs insertion validation and characterization: 1) 3′ junction PCR (3′PCR) with one primer specific to L1HS (L1Hs-AC-22) and the L1-IP 5′ peak flank primer (upstream to the peak) designed above, was used to verify the presence of the predicted insertion; and 2) long-range full-length PCR (FL-PCR) with the L1-IP 5′ and 3′ peak flank primers designed above, was used to clone the entire length of candidate insertions and also to determine the zygosity (homozygous versus heterozygous) of insertions. All PCR products were run on 1% agarose gels, and images were analyzed by ImageQuant TL software (GE Healthcare) to quantify the product sizes, relative intensities and absolute peak heights of the bands in an unbiased manner.

Previously published PCR protocols for L1Hs validation (Ewing and Kazazian, 2010; Iskow et al., 2010; Stewart et al., 2011) were adapted and optimized to maximize sensitivity and specificity for both unamplified bulk and MDA-amplified single-cell DNA. 3′PCR and FL-PCR protocols are shown in Tables S6A and S6B, respectively.

3′PCR. Positive 3′PCR reactions yield a PCR product within 150bp size of the predicted size. The predicted size was calculated as the distance from the 5′ peak flank primer to the 3′ end of the L1-IP peak plus 114bp, which is the distance of the AC primer location to the end of the L1Hs consensus sequence, plus 50bp which is the approximate expected poly-A tail length for recent polymorphic and disease-causing insertions (Beck et al., 2010; Hancks and Kazazian, 2012). Peak coordinates do not necessarily precisely border the insertion in situations where seed sequences are not present adjacent to the insertion, low mappability may prevent read mapping adjacent to the insertion, and poly-A tail lengths are variable, which leads to only approximate predicted sizes. The difference between observed to predicted 3′PCR product sizes was -8 ± 39 bp (SD) for 71 out of 76 insertions that validated in the 3′ PCR validation screen (see below), supporting our ability to predict amplicon size within 150bp. Negative reactions yield no PCR product or in rare cases a band outside of the predicted size range.

FL-PCR. For heterozygous insertions genotyped by FL-PCR, two products are expected: a smaller product within 150bp of the predicted size without an L1Hs insertion, and a product ranging up to ~6kb larger than the smaller product, depending on the size of the L1Hs insertion.

3′PCR Sensitivity and Specificity Calculations

The sensitivity and specificity of the 3′PCR validation method was assessed by performing validation of 64 high-confidence known germline insertions found by L1-IP in bulk tissues (33 KR and 31 KNR). In a separate experiment, 3′PCR validation was performed on 12 high-confidence unknown (UNK) germline candidate insertions found by L1-IP in bulk tissues. Among these 76 insertions (64 known and 12 unknown), 31 were present in all 3 individuals in this study and 45 were absent from at least 1 of the 3 individuals (polymorphic) (see Table S3).

Sensitivity. The sensitivity of 3′PCR was 92% (70/76) with unamplified bulk DNA, 92% (70/76) with MDA-amplified unsorted (50×10^3 cells), and 93% (71/76) with MDA-amplified single-cell (one sample of each type assayed), demonstrating consistent sensitivity of 3′PCR for both unamplified and MDA-amplified DNA (Figure S6E). 2/5 of the insertions that failed validation did not have any visible PCR product, and 3/5 had product of the wrong size. The PCR sensitivities for the KR, KNR, and UNK insertions were 91% (30/33), 94% (29/31), and 92% (11/12), respectively, in both unamplified bulk DNA and unsorted 50,000-nuclei amplified DNA, and 91% (30/33), 97% (30/31), and 92% (11/12), respectively, in 1-neuron amplified DNA. For candidates that failed the initial 3′PCR validation, up to 4 additional 5′ peak flank primers were designed and tested to differentiate PCR failure from false insertion predictions of the L1-IP pipeline. The PCR sensitivity after an additional second set of primers increased to 97%, 97% and 96% in bulk DNA, unsorted nuclei and 1-cell, respectively. After testing failed candidates on up to four sets of primers, the final sensitivities were 100% for bulk DNA, unsorted-nuclei and 1-cell samples, confirming that the initial loss of sensitivity was due to faulty primers.

Specificity. The specificity of the 3′PCR method for loci without predicted insertions was also determined by assaying for polymorphic L1Hs loci in individuals predicted not to have an insertion by L1-IP. These loci should not yield a band of the predicted size. The 45 polymorphic germline insertions were assayed in the individuals predicted by L1-IP to not have the insertions. The experiment was performed on one sample each of bulk DNA, amplified unsorted-nuclei and single-cell, from each of the individuals without the insertion. No false positive validations were observed in bulk DNA, unamplified-nuclei, and 1-cell samples (63 reactions each). 187/189 of the reactions had no band. 2/189 had a band > 1kb larger than the predicted size in bulk and 1-cell samples of one individual, and

failed validation for this reason. The specificity of the 3′PCR for loci without a predicted insertion (i.e., no L1-IP peak) was therefore 100%.

L1Hs 3′PCR Single-Cell Allelic and Locus Dropout Rates

A comprehensive assessment of AD and LD of L1Hs insertions was carried out by 3′PCR genotyping of 3 heterozygous (for AD estimation) and 3 homozygous (for LD estimation) insertions in each of 83 single neurons from individual 1465. 'Low amplification' in 3′PCR was defined as callable alleles with peak height < 10,000 arbitrary fluorescence units. Peaks with height < 5,000 units were considered as 'allelic dropout' as these are barely above background noise. Peak height with ≥ 10,000 arbitrary fluorescence units were considered 'normal amplification'.

TOPO-TA Cloning and Sanger Sequencing

PCR products were sequenced either by direct Sanger sequencing of PCR products, or by TOPO-TA cloning (Life Tech.) for subsequent Sanger sequencing. All Sanger sequencing was performed by Genewiz. Sequence traces were analyzed and assembled by Geneious (Biomatters Ltd.).

Cloning of False Positive Chimeras

Candidate false positive chimeras, adjacent (≤20kb) to known L1Hs or L1Pa insertion peaks found in the L1-IP data, can be differentiated from true insertions if cloning and sequencing of the chimera shows one or more of the following features: a) a breakpoint within the 3′UTR of the L1, leading to absence of a polyA tail; b) microhomology between the adjacent known L1 insertion and the chimera at the breakpoint site; c) sequence from genomic DNA downstream of the adjacent known L1 in the cloned chimera fragment; d) sequence from genomic DNA upstream of the adjacent known L1 in the cloned chimera fragment. False positive chimera candidates were first screened by 3′PCR as described above, on both the MDA-amplified sample predicted to contain the chimera and on unamplified bulk DNA from the individual as a negative control. PCR product was obtained for candidate chimera A and was TOPO-cloned and Sanger sequenced (Figure S5). FL-PCR was also performed to determine the sequence upstream of the chimeric L1. See Table S3 for a list of the chimeras tested.

AKT3 Mosaic Mutation Sequencing Analysis

HMG-3 resected malformed cortical tissue was FACS sorted as described above, or used to extract bulk DNA, from one of two locations ~5cm apart. $NeuN^+$ and $NeuN^-$ cells were sorted as above, and large and wide nuclei were sorted using FSC gates to determine whether larger nuclei are enriched for the *AKT3* mutation which is known to lead to abnormally large cells and nuclei. The *AKT3* c.49G→A (E17K) mutant locus was amplified by PCR as previously described (Poduri et al., 2012), with primers listed in Table S3. Sanger sequencing with the forward PCR primer had negligible background signal at the mutant locus in HMG-3 blood and normal samples and allowed quantification of the mutant allele frequency in 10-, 100-, 1,000-, 10,000-cell, and bulk samples by comparing normal to mutant allele peak intensity in the Sanger traces with the Mutation Surveyor software's (SoftGenetics) mosaic mutation quantifier function and correcting for background noise (1.6%) in tissue from 8 unrelated normal controls. Mosaicism in bulk DNA, 1,000-$NeuN^+$-nuclei, and 1,000-$NeuN^-$-nuclei samples was further assessed by cloning PCR products with the TOPO TA kit (Life tech.) and sequencing to count the number of clones harboring the mutant allele. In Sanger trace quantification and TOPO cloning experiments, the percent mosaicism (% of mutant cells in the tissue) was calculated as 2 × % mutant allele since the mutant allele is heterozygous (see verification of the heterozygous state below). In single-cell experiments, since single-cell isolation 'clones' each cell's genome, mosaicism is quantified by directly counting the number of cells harboring the mutation, regardless of the mutant versus normal allele peak intensity. Therefore, any detection of the mutant allele above background levels is counted as a mutant cell. The allelic dropout (AD) rate (leading to false negative mutant cells with only the normal allele amplified) was calculated as (# of cells with only the mutant allele, i.e., undetectable normal allele) / (2 × # of cells with only the mutant allele + # of cells with both normal and mutant alleles). The AD measured in this experiment was 20%. To correct for AD, the percent mosaicism was estimated as (# of cells with both the mutant and normal alleles + 2 × # of cells with only the mutant allele) / (total # of cells). The standard error of the percent mosaicism measured from Sanger trace allele intensities was estimated as the standard deviation of replicates when replicates were available, since these correspond to binomial sampling experiments where the sample size is very large and error derives mainly from error in the Sanger sequencing rather than sampling error. The standard error of the percent mosaicism measured in single-cell experiments was estimated as $\sqrt{p \cdot (1-p)/n}$, or twice this value for TOPO experiments, where p is the fraction of mutant cells or clones out of n cells or clones assayed.

Determining the zygosity of mosaic mutations is only possible with a single-cell analysis, since bulk DNA sequencing of mosaic mutations is blind to the fraction of cells in the tissue with the mutant allele. We therefore empirically determined the zygosity of the mutant *AKT3* locus by analyzing the sequence data from the locus in all single amplified nuclei. Seventy-four percent of cells with the mutant allele also had a detectable normal allele (corresponding to 20% AD of the normal allele) indicating that both the normal and mutant *AKT3* alleles are present in mutant cells, excluding a hemizygous mutant genotype. Next, by quantifying the relative signal of the normal and mutant alleles in single cells, we determined whether the mutant and normal alleles were present in equal copy numbers to exclude the possibility of selection for a lineage with amplification of the mutant allele. The average mutant allele peak height fraction across all single cells harboring both the mutant and normal alleles was 44% ± 4% (SEM), consistent with a normal allele and a mutant heterozygous dominant activating mutation present in equal copy numbers.

SUPPLEMENTAL REFERENCES

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). Isolation of an active human transposable element. Science *254*, 1805–1808.

Esteban, J.A., Salas, M., and Blanco, L. (1993). Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. J. Biol. Chem. *268*, 2719–2726.

Ewing, A.D., and Kazazian, H.H., Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. *21*, 985–990.

Goecks, J., Nekrutenko, A., and Taylor, J.; Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11*, R86.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T., et al. (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171–1182.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lyck, L., Dalmau, I., Chemnitz, J., Finsen, B., and Schrøder, H.D. (2008). Immunohistochemical markers for quantitative studies of neurons and glia in human neocortex. J. Histochem. Cytochem. *56*, 201–221.

Nielsen, S., Nagelhus, E.A., Amiry-Moghaddam, M., Bourque, C., Agre, P., and Ottersen, O.P. (1997). Specialized membrane domains for water transport in glial cells: high-resolution immunogold cytochemistry of aquaporin-4 in rat brain. J. Neurosci. *17*, 171–180.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Development Core Team (2011). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna. http://www.r-project.org.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. *132*, 365–386.

Smit, A.F.A., Hubley, R., and Green, P. (2010). RepeatMasker Open-3.0. http://www.repeatmasker.org.

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.-P., et al.; 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. *7*, e1002236.

Sudhof, T.C. (2004). The synaptic vesicle cycle. Annu. Rev. Neurosci. *27*, 509–547.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., and Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum. Mutat. *27*, 323–329.

Yamada, T., Soma, H., and Morishita, S. (2006). PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome. Nucleic Acids Res. *34*, W665–W669.

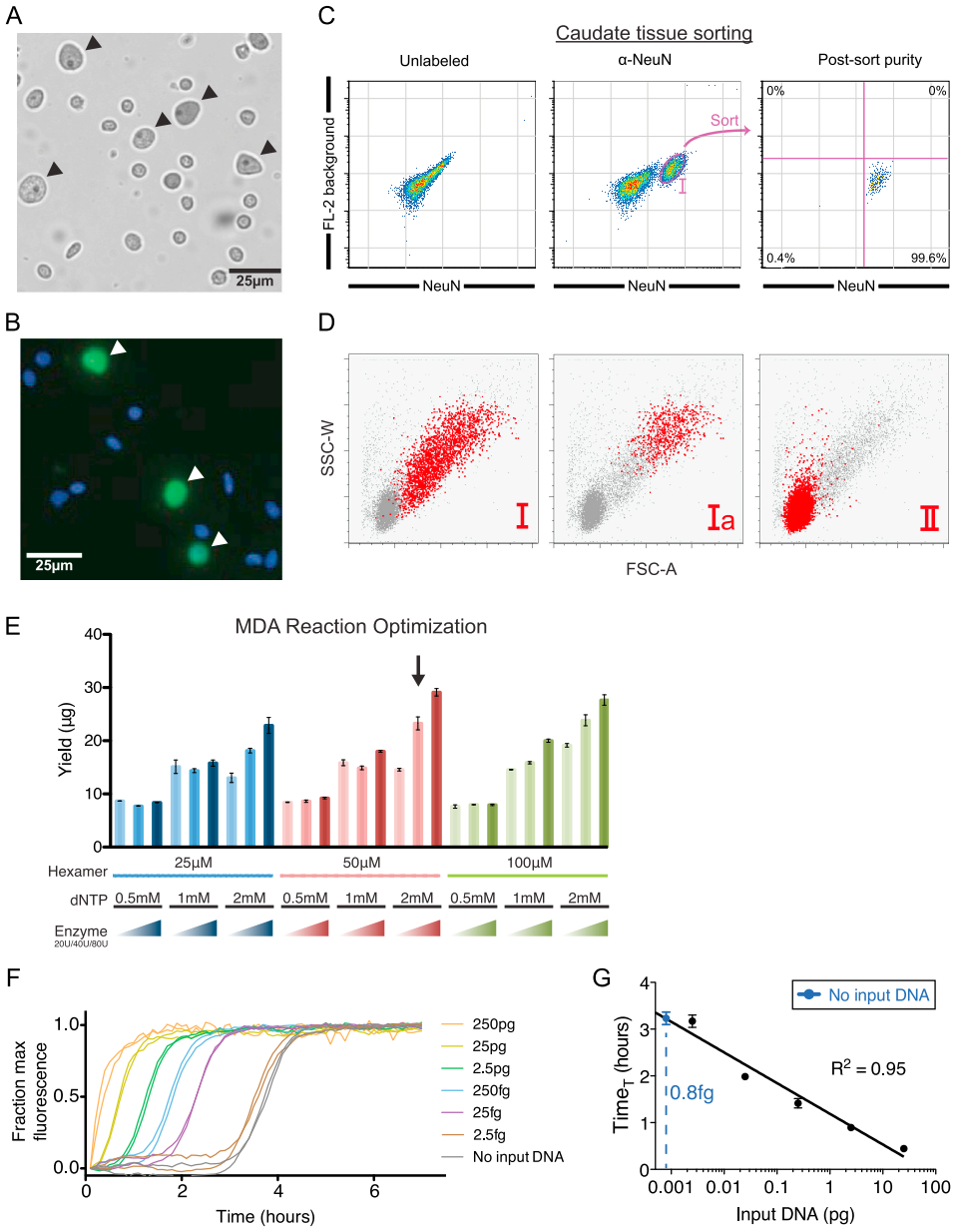

Figure S1. Nuclei Purification, Sorting of Caudate Tissue, and MDA Reaction Optimization, Related to Figure 1

(A) Purified nuclei from postmortem human frontal cortex. Nuclei with neuronal nuclei morphology (large, prominent nucleolus) can be readily observed (arrowheads). Pyramidal shape in some large nuclei is reminiscent of pyramidal neuronal nuclei shape.

(B) NeuN (green) and Hoechst (blue) staining of cortical nuclei by fluorescence microscopy. White arrowheads indicate neuronal nuclei.

(C) FACS of NeuN-labeled caudate nuclei.

(D) Forward (FSC) and side (SSC) scatter backgating analysis of flow cytometry of human cortical nuclei (Figure 1B). Red dots indicate events from the specified population (I, Ia, and II) out of all events recorded (gray dots). Population I nuclei have a distinct size distribution larger than population II, and population Ia nuclei are larger nuclei within population I.

(E) MDA reaction yields of control purified human DNA amplified with varying hexamer, dNTP and phi29 polymerase concentrations (error bars ± 1SD, n = 2 per condition). Arrow indicates chosen reaction conditions for single-neuron genome amplification. Yields from single-neuron genome amplification (15-20 μg) are slightly less than yields from control human DNA with the chosen reaction conditions.

(F) Real-time quantitative MDA monitoring of amplification reactions with varying lambda DNA input.

(G) A semi-log standard curve was fit to data from amplification curves in Figure S1F (Time$_T$, time-to-threshold-amplification). No-input DNA reactions have ~0.8 femtograms of non-human exogenous DNA, negligible compared to 6.5 picograms of DNA in a single human nucleus (error bars ± 1SD, n = 2 per condition).

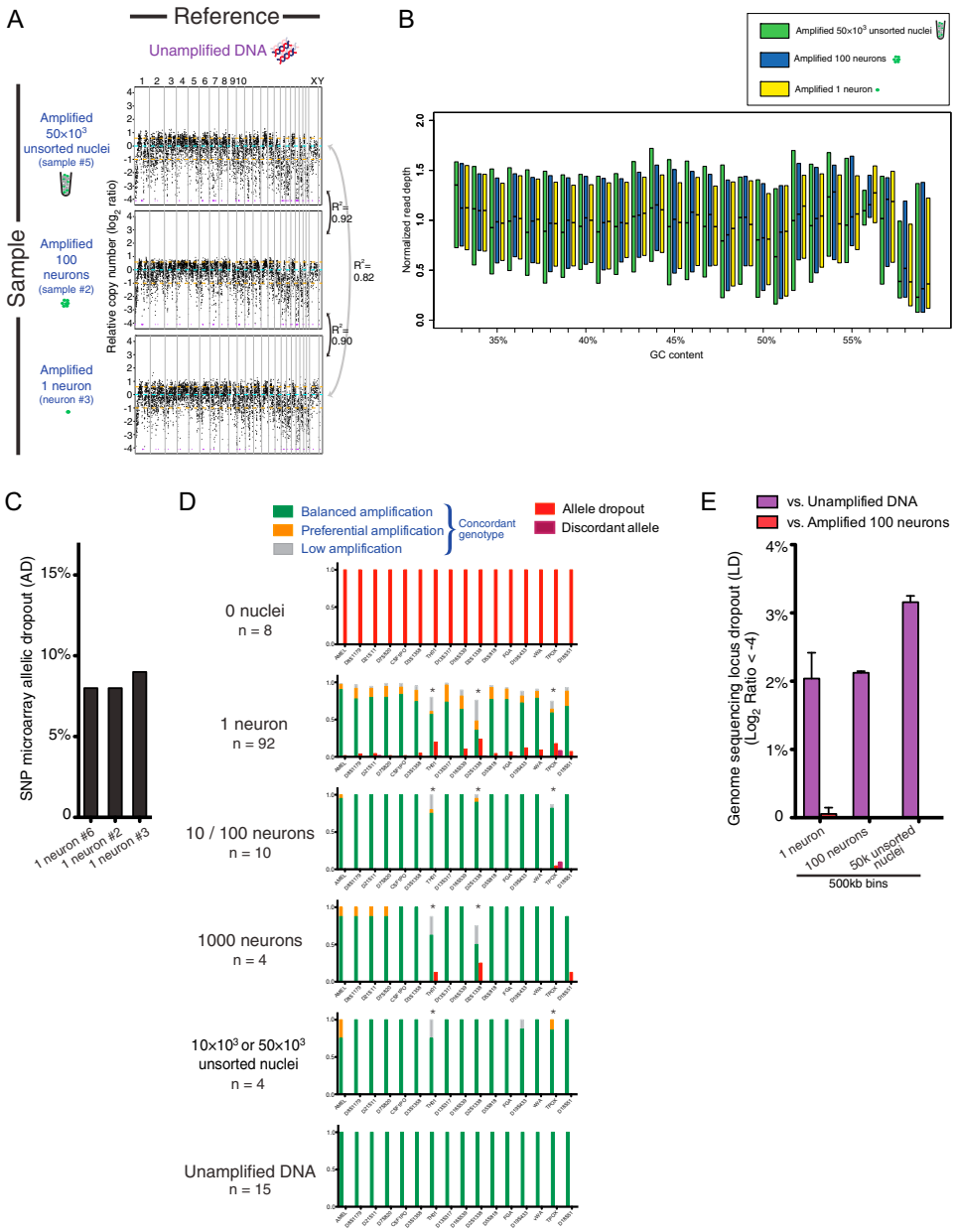

Figure S2. Low-Coverage Sequencing and Identifiler Assessment of Single Neuronal Genome Amplification Quality, Related to Figures 2A–2D

(A) Copy number profiling in 6,000 equal-read bins of ~500 kb in size, relative to an unamplified DNA reference, shows that MDA bias is consistent and reproducible regardless of the number of nuclei amplified. Correlations of bin copy numbers (R^2) between 50 × 10^3-nuclei, 100- and 1-neuron samples are shown. Purple points represent off-scale bins.

(B) Genome coverage from low-coverage whole-genome sequencing experiments was calculated for 5kb non-overlapping windows and binned by GC content. Read depth for each sample was normalized to the average read depth across all bins. Boxplots show median and upper and lower quartile boundaries.

(C) Allelic dropout (AD) in single neurons measured by SNP microarray genotyping.

(D) Per-locus Identifiler results for different sample types. Asterisks mark consistently under-performing loci, indicating that they reside in regions that do not amplify well by MDA.

(E) Genomic locus dropout (LD) estimates from low-coverage sequencing, normalized to the indicated unamplified and amplified references (error bars ± SD, n = 4 for 50 × 10^3 nuclei, n = 2 for 100-neuron, and n = 6 for 1-neuron groups). Locus dropout was estimated as the percentage of low-coverage sequencing bins with less than 1/16 the copy number relative to the reference ($\log_2$ ratio < −4). Using 100-neuron samples as a reference, LD in 1-neuron samples was lower at 0.05%, again consistent with the finding that most regional amplification bias is inherent to MDA.

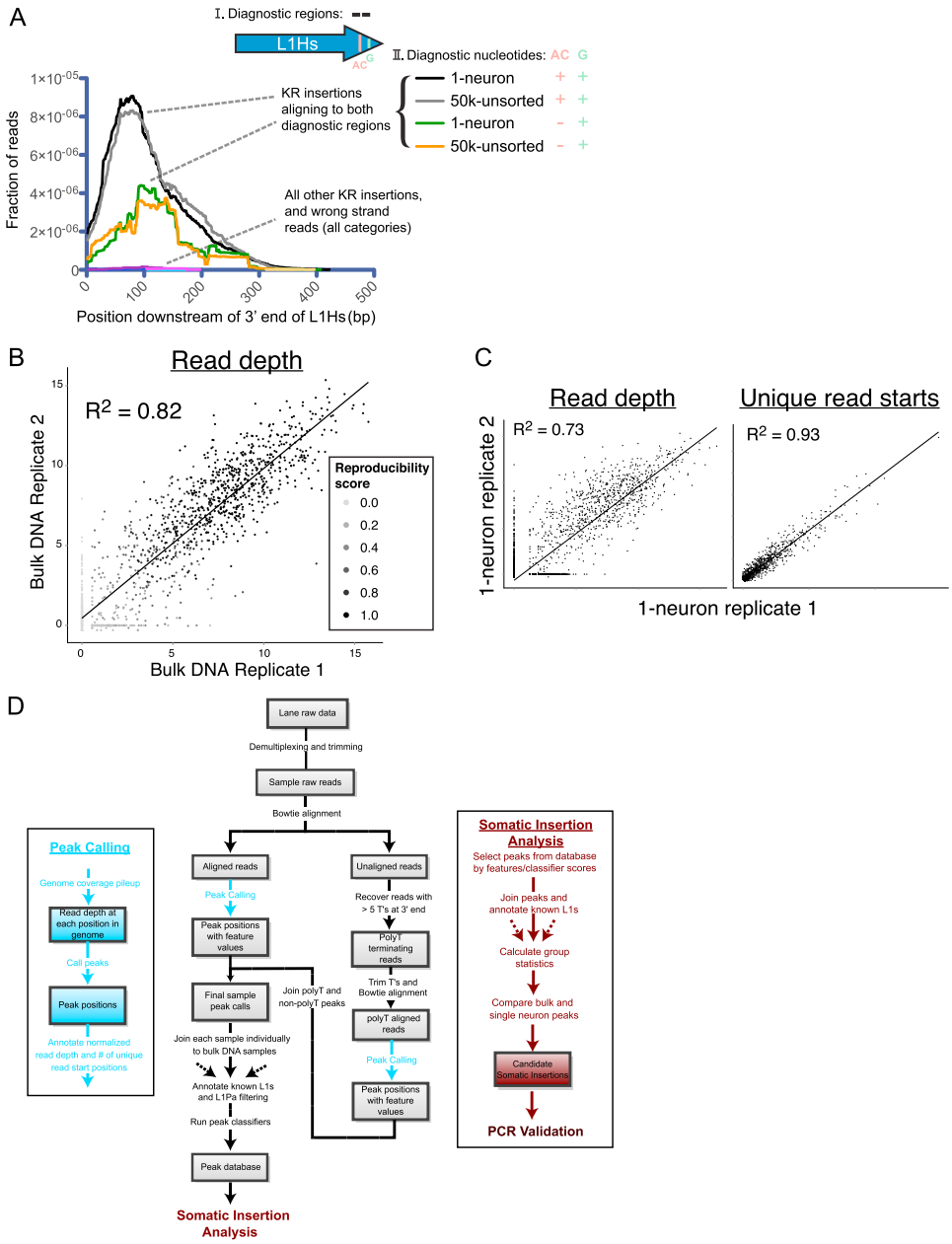

Figure S3. L1-IP Specificity and Reproducibility and the Computational Pipeline for Somatic Insertion Analysis, Related to Figure 3

(A) L1Hs AC and G diagnostic nucleotides are discriminated by L1-IP. Histogram of L1-IP library sequencing read coverage downstream and pointing toward the 3′ end of all known reference (KR) L1Hs insertions. Each KR insertion is categorized according to whether it aligns to one, neither, or both L1Hs diagnostic regions, as described in Extended Experimental Procedures and Table S5. Additionally, KR insertions are annotated according to whether the AC or G diagnostic bases are present within the aligned regions. KR insertions aligning to both diagnostic regions, with both the AC and G diagnostic bases present (and to a lesser extent those with only the G diagnostic base), are detected by L1-IP. Also graphed is the summed coverage of all reads pointing away from the 3′ end (wrong strand reads) regardless of category. Data shown is an average for n = 8 unsorted 50,000-nuclei and n = 30 1-neuron samples. Coverage is normalized to the number of insertions in each category and the total number of reads in the samples.

(B) Correlation plots of normalized read depth ($\log_2$ transformed) for all peaks in one representative pair of technical replicate L1-IP libraries from bulk DNA cortex of individual 1465. Points are shaded by the reproducibility scores of the peaks (see Extended Experimental Procedures), and graphed is a linear regression fit to the data.

(C) Correlation plots of normalized read depth ($\log_2$ transformed) and the number of unique read start sites for all peaks in a representative pair of technical replicate L1-IP libraries from 1 cortical neuron of individual 1465. Similar results were obtained for unsorted 50,000-nuclei and 100-neuron technical replicates (data not shown).

(D) Schematic of the L1-IP data analysis pipeline; full details are provided in Extended Experimental Procedures.

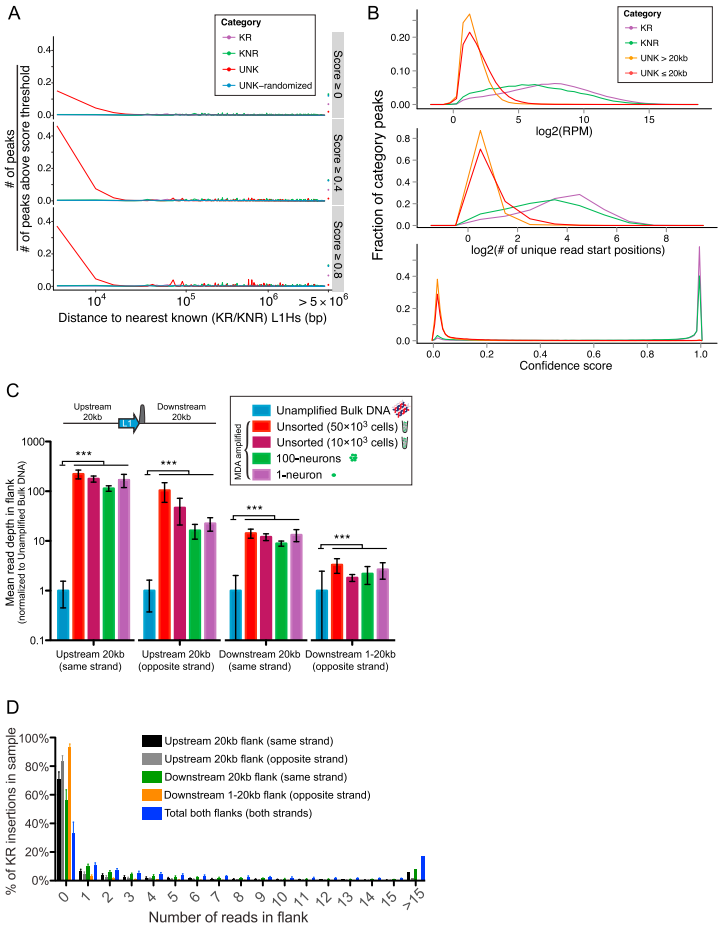

Figure S4. Characterization of MDA Chimera Peaks and Reads, Related to Figure 3

(A) Histogram of the distance of peaks in different categories to the nearest known (KR or KNR) L1Hs insertion detected in bulk samples, for all peaks in all single-cell samples (n = 303 samples). Genomic positions of UNK peaks were also randomized (UNK-randomized) to determine the distribution expected by random unbiased insertion in the genome. Results are shown for different confidence score thresholds, with the *y* axis indicating the number of peaks at each distance divided by the number of peaks above the given score threshold (i.e., the fraction of peaks above the score threshold). The percentage of peaks < 20 kb from the nearest known L1Hs (with a score threshold of 0.4) was 60.4% for UNK peaks, as opposed to 0.7%, 1.3%, and 1.1% for KR, KNR, and UNK-randomized peaks, respectively.

(B) Histogram of normalized read depth (RPM), number of unique read start positions, and confidence scores for KR, KNR, and UNK peaks in all single-neuron samples. UNK peaks are separated into peaks located less than 20kb and farther than 20kb from the nearest known (KR/KNR) L1Hs. KR and KNR peaks have significantly higher read depth, number of unique read start positions and confidence scores than unknown peaks.

(C) Mean read depth (±SD), normalized to bulk DNA, in upstream and downstream 20kb flanks of detectable KR L1Hs insertions (i.e., containing L1Hs diagnostic regions targeted by L1-IP), shown separately for each sample type for all samples in the study (n = 31 bulk DNA, n = 13 unsorted 50,000-nuclei, n = 5 unsorted 10,000-nuclei, n = 15 100-neuron, and n = 303 1-neuron samples). Insertions containing other KR, KNR or L1-IP-filtered L1Pa insertions within 50 kb were excluded from the analysis to avoid counting non-chimera reads mapping to those peaks, leaving 612 peaks for the analysis. Reads on the same and opposite strands as the L1Hs were counted separately. For the downstream opposite strand category, 0 to 1kb downstream of the L1 was excluded to avoid counting the L1-IP peak reads. There are increased reads in MDA-amplified samples on both strands upstream and downstream of insertions compared to unamplified bulk DNA (t test, $p < 10^{-15}$). Flank read depth differences between 1-neuron and unamplified bulk samples were 169-fold, 23-fold, 13-fold, and 3-fold for upstream-same, upstream-opposite, downstream-same, and downstream-opposite reads, respectively. Same-strand reads correspond to inversion chimeras, consistent with a branch migration mechanism (Lasken and Stockwell, 2007). The larger increase in upstream-opposite versus downstream-opposite strand reads in MDA samples may be due to blocking of downstream DNA with a downstream amplicon, while amplicons can rehybridize to the upstream template DNA they were previously hybridized to, before a new amplicon is synthesized. There was also a trend for decreasing flank chimera reads with fewer cells input into the MDA reaction, likely due to fewer cross-priming chimera events when less initial DNA is present in the same MDA reaction volume. 1-neuron samples, however, have slightly more flank chimera reads than 100-neuron samples, likely because any unique MDA chimera event formed early in the amplification in a 100-neuron sample comprises a smaller fraction of the total DNA, than in a 1-neuron sample. While chimera reads are present in MDA-amplified samples, they comprise only a small fraction of the total sample reads: on average 0.5% for unsorted 50,000-nuclei MDA, 0.3% for unsorted 10,000- nuclei MDA, 0.2% for 100-neuron-MDA, and 0.3% for 1-neuron-MDA samples.

(D) The percentage of KR insertions detected in each 1-neuron sample (mean for n = 303 1-neuron samples, ± SD) that have the specified number of chimera reads in the local 20 kb flanks. Results are shown separately for upstream and downstream flanks, and for reads in the same and opposite orientation to the insertion. KR insertions included in the analysis were the same as in Figure S4C.

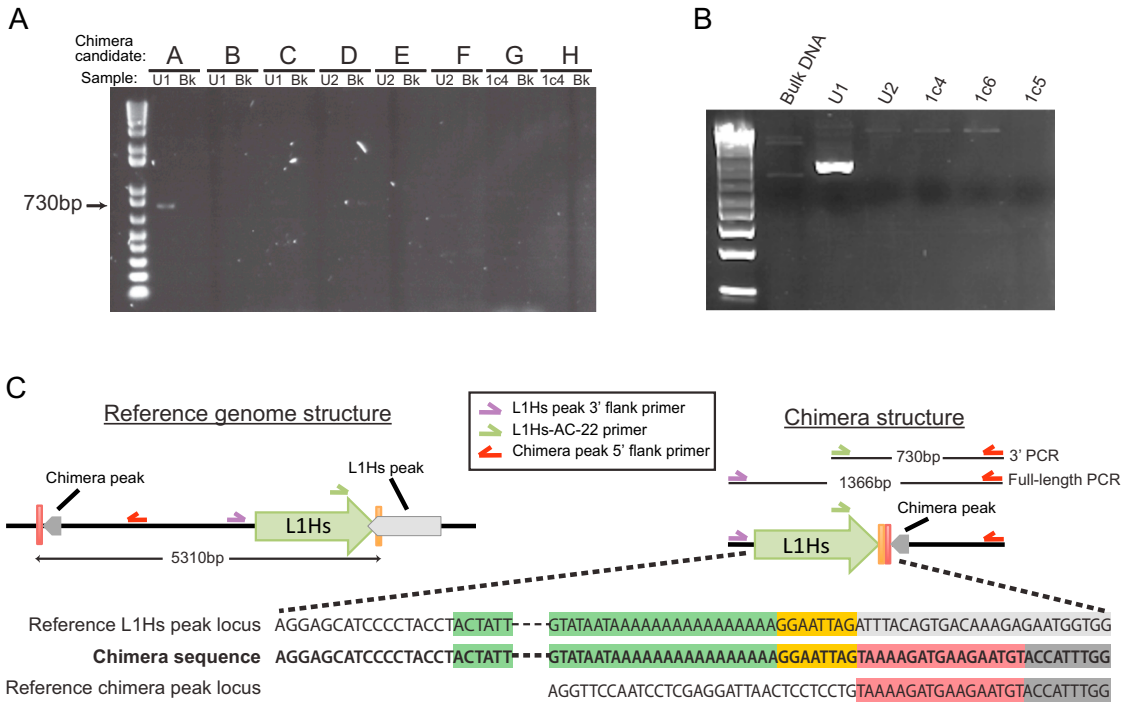

Figure S5. MDA Chimera Screening and Cloning, Related to Figure 3

(A) 3′PCR screen for 8 chimera candidates from 3 different MDA amplified samples. 'U1', 'U2', and '1c4' represent the samples: unsorted 50,000-nuclei #1, unsorted 50,000-nuclei #2, and 1-neuron #4, respectively, from the cortex of individual 1465. 'Bk' represents reactions performed with unamplified bulk control DNA from the same individual. Only chimera candidate A, adjacent to L1Hs-KR-chr3_121120147, yielded a detectable 3′PCR product.

(B) 3′PCR of chimera candidate A in 6 different samples. As in Figure S5A, PCR product was only detected in MDA-amplified unsorted #1 (U1). This indicates that each chimera event is unique to a given sample. Sample name abbreviations are as in Figure S5A.

(C) Structure of chimera A. Upper left panel: schematic of the human genome reference at the location of chimera A and its adjacent known L1Hs (L1Hs-KR-chr3_121120147). Upper right panel: schematic of the chimera structure. Bottom panel: sequence of the cloned chimera. The green arrow represents the adjacent known L1Hs from which the chimera derived; the light and dark gray arrows represent the L1Hs and chimera peaks, respectively, detected by L1-IP; the orange box is the sequence downstream of the L1Hs insertion up to the breakpoint; and the red box is the sequence downstream of the chimera peak up to the breakpoint. 3′PCR was performed using the L1Hs specific primer (L1Hs-AC-22; green) and the peak 5′ flank primer of the chimera (red). In addition to 3′PCR characterization, we cloned the full chimera by full-length (FL)-PCR using the L1Hs peak 3′ flank primer (purple) with the chimera peak 5′ flank primer (red). Product from this FL-PCR would not be produced in the absence of such a chimera since the primers would point away from each other.

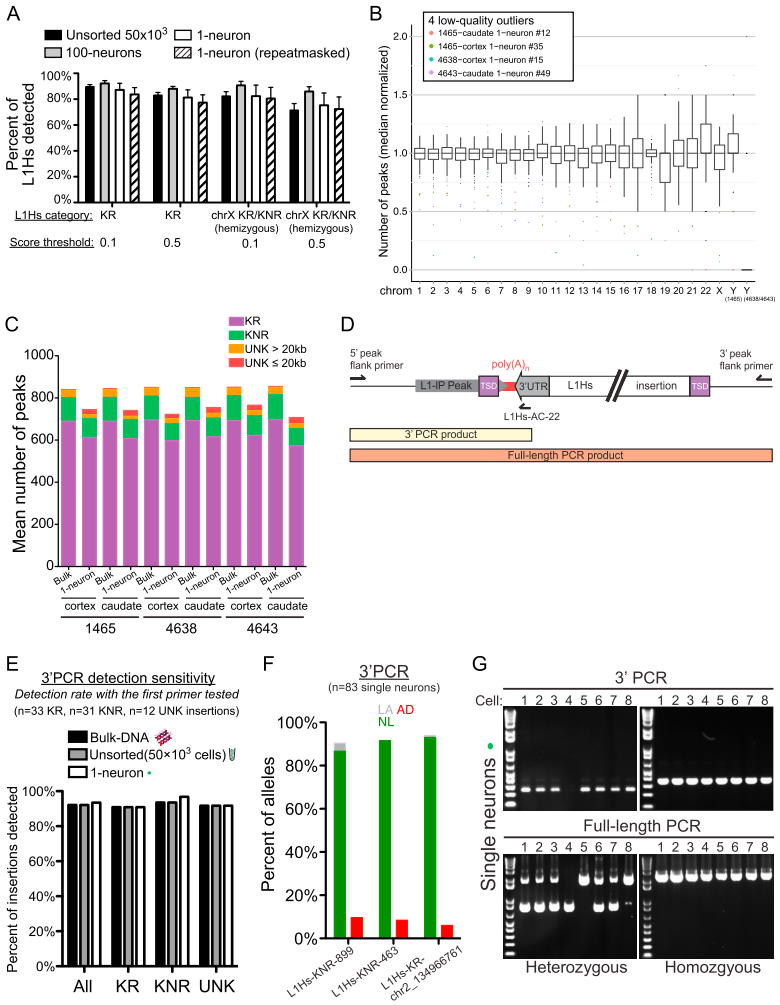

Figure S6. L1-IP Sensitivity and PCR Validation in Single Neurons, Related to Figure 3

(A) Mean (±1 SD) L1-IP sensitivity to detect germline L1Hs insertions in unsorted 50,000-nuclei, 100-neuron, 1-neuron samples, as well as 1-neuron samples mapped to an L1Hs repeat-masked genome (to control for the possibility of increased detection of KR insertions present in the genome reference). Sensitivity is calculated for known reference (KR) insertions across all unsorted 50,000-nuclei, 100-neuron and 1-neuron samples in the study, and for chrX KR and KNR hemizygous insertions across all unsorted 50,000-nuclei, 100-neuron and 1-neuron samples of individual 1465 (male). Germline insertions are defined as insertions found in at least half of the bulk DNA samples of the individual (i.e., at least 3 tissues), and with a minimum confidence score of 0.5 in at least one of these. Sensitivity is calculated for confidence score thresholds of 0.1 or 0.5.

(B) Box plot of number of L1-IP peaks, with confidence score > 0.5, per chromosome, for all 1-neuron samples in the study. Outliers are represented by black dots. 4 consistent outlier low-quality samples are colored, which include the 3 low-quality samples clustering separately in Figure 5A.

(C) Mean number of peaks with score > 0.5 across all bulk and 1-neuron samples from cortex and caudate of individuals 1465, 4638 and 4643. The number of peaks is shown separately for different categories (KR, KNR, UNK peaks at least 20 kb from the nearest KR or KNR insertion found in bulk DNA samples, and UNK peaks within 20 kb of the nearest KR or KNR insertion).

(D) Schematic of the PCR validation methods. Two primers are designed flanking each L1-IP peak, and two types of validation can be performed: 3′ junction PCR (3′PCR) to confirm the existence of a predicted insertion, and full-length PCR (FL-PCR) for cloning of the complete insertion.

(E) 3′PCR detection rate (3′PCR sensitivity), with the first 5′ peak flank primer tested, of 33 KR, 31 KNR, and 12 unknown (UNK) germline insertions detected with high confidence by L1-IP. 3′PCR was performed on bulk DNA, unsorted 50,000-nuclei, and 1-neuron samples (one sample of each type tested). Sensitivity of 92% in bulk DNA and unsorted 50,000-nuclei samples and 93% in 1-neuron samples indicates a high technical sensitivity and consistent performance of 3′PCR validation for unamplified and MDA-amplified DNA. All insertions that were not detected with the first primer were subsequently detected successfully with redesigned primers.

(F) 3′PCR quantification of AD of 3 heterozygous L1Hs insertions in 83 single neurons from individual 1465 cortex. Combined AD rate for all 3 insertions is shown in Figure 3E. NL, normal amplification; LA, low amplification; AD, allelic-dropout.

(G) Representative gel images of 3′PCR and full-length PCR of 1 homozygous and 1 heterozygous L1Hs germline insertion in 8 different single neurons. The upper band in full-length PCR of the heterozygous insertion is the allele with the transposon insertion, and the lower band is the allele with no transposon insertion. Although the majority of cells have both alleles evenly amplified, AD of the insertion allele can be seen in some cells (e.g., neuron #4), which correlates with absence of 3′PCR product in the same cell. AD of the allele with no insertion can be seen in neuron #5. Neuron #8 had preferential amplification of the insertion allele.

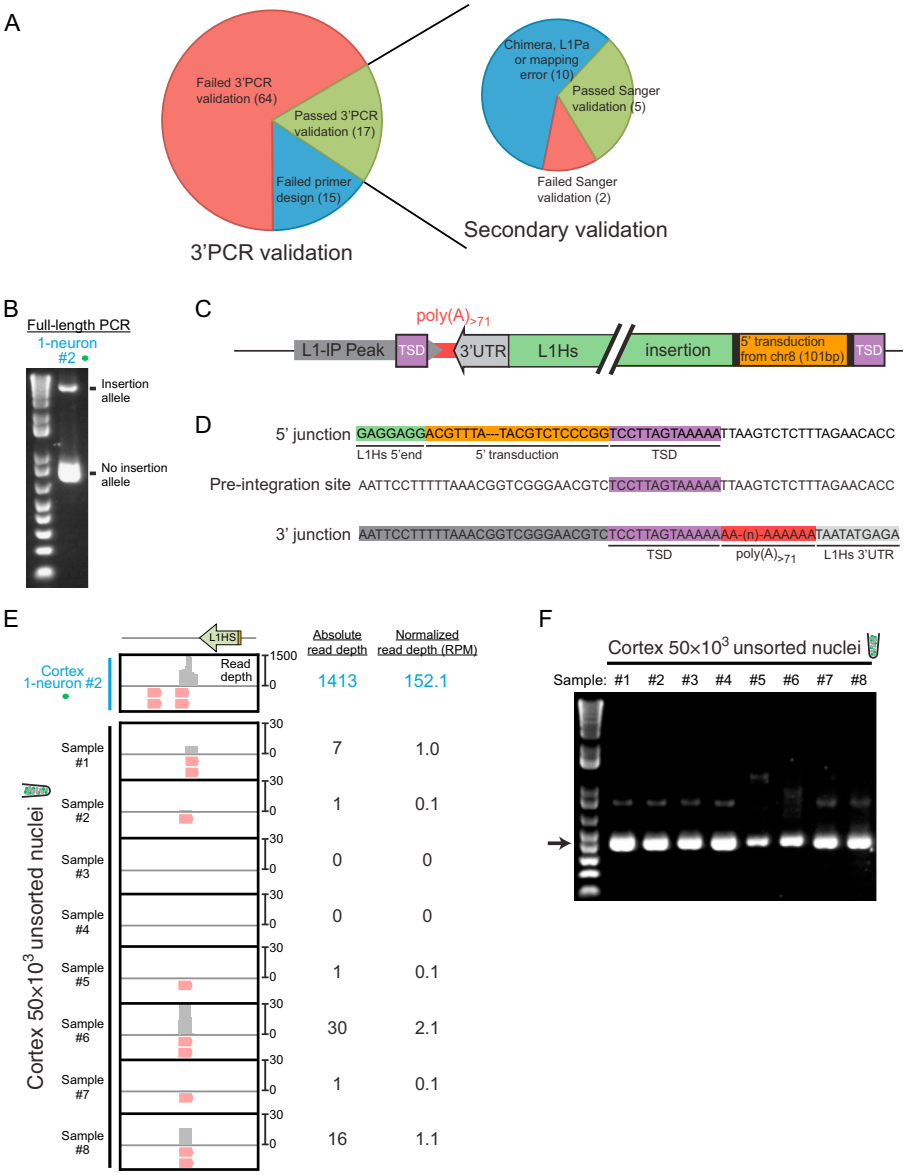

Figure S7. Validation of L1-IP Somatic L1Hs Insertion Candidates, Related to Figure 6

(A) 3′PCR and secondary validation summary of top-scoring somatic insertion candidates identified by L1-IP.

(B) Gel image of full-length (long-range) PCR cloning of the validated L1Hs somatic insertion shown in Figures 6C-E (L1-IP peak ID chr15_67625710_plus_0_0), in sample 1465-cortex 1-neuron #2.

(C) Structure of the L1Hs somatic insertion (chr15_67625710_plus_0_0), cloned by full-length PCR in Figure S7B. Pre-integration TSD coordinates are chr15: 67,625,702-67,625,714 (hg19). 5′ transduction (chr8: 73,793,824-73,794,027) from upstream of the source L1Hs (L1Hs-KR-chr8_73787792) exhibited transcriptional splicing removing 103bp (chr8: 73,793,831-73,793,934) from the source sequence. The source L1Hs on chr8 is a full-length, intact insertion that is polymorphic in the population. Details of junction sequences are in Figure S7D.

(D) Sequences of 5′ and 3′ junctions of the L1Hs somatic insertion shown in Figures 6C-E and S7C (chr15_67625710_plus_0_0). Poly-A tail length (at least 71bp long) was not possible to determine exactly due to difficulty sequencing through homopolymeric regions. The full sequence of the insertion, including the 5′ transduction and flanking genomic sequences, is in Table S3.

(E) L1-IP detects the somatic L1Hs insertion shown in Figures 6C-E (chr15_67625710_plus_0_0) at low levels in some unsorted 50,000-nuclei samples from the cortex of individual 1465, as expected for a low-level mosaic insertion. Absolute and normalized (reads per million mapped reads, RPM) read depths are shown. A maximum of two raw reads per position are shown below the read depth histogram tracks. The eight unsorted 50,000-nuclei samples were taken from the same piece of tissue from which 1-neuron #2 and 1-neuron #77 derived. Reads at this somatic insertion locus were not seen in other bulk samples from individual 1465 (data not shown).

(F) Optimized 3′PCR, with increased DNA input and higher-cycle PCR, detects the somatic L1Hs insertion shown in Figures 6C-E (chr15_67625710_plus_0_0) in all eight unsorted 50,000 cortical nuclei samples from individual 1465. 3′PCR product sequencing from one of the eight samples confirmed the correct sequence of the somatic insertion.