Evidence that *APP* gene copy number changes reflect recombinant vector contamination

Junho Kim[1], Boxun Zhao[1], August Yue Huang[1], Michael B. Miller[1,2,3], Michael A. Lodato[1,2], Christopher A. Walsh[1,2]* and Eunjung Alice Lee[1]*

[1]Division of Genetics and Genomics, Manton Center for Orphan Disease, Boston Children's Hospital, Boston, MA, USA; Department of Pediatrics, Harvard Medical School, Boston, MA, USA; and Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[2]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA and Department of Neurology, Harvard Medical School, Boston, MA, USA

[3]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

*to whom correspondence should be addressed. E-mail: Christopher.Walsh@childrens.harvard.edu or EAlice.Lee@childrens.harvard.edu

Abstract

Lee et al.[1] (hereafter "the Lee study") have recently reported that RNA-mediated somatic recombination or somatic retrotransposition of the *APP* gene occurs in neurons from both control individuals and those with sporadic Alzheimer's disease (AD). As evidence of somatic *APP* retrotransposition, the authors present various forms of *APP* genomic cDNA (gencDNA) in PCR-based (Sanger sequencing, SMRT sequencing) and non-PCR-based (targeted hybrid-capture sequencing, DNA in situ hybridization (DISH)) experiments. They also report greater prevalence of *APP* gencDNA in AD neurons compared to control neurons (69% *vs* 25% of neurons with at least one *APP* retrogene insertion on average, Fig. 5 and Extended Data Fig. 5 in the Lee study) as well as its greater diversity. We reanalyzed the *APP*-targeted sequencing data from the Lee study, revealing evidence that *APP* gencDNA originates mainly from the contamination by exogenous *APP* recombinant vectors, rather from true somatic retrotransposition of endogenous *APP*. We also present our own single-cell whole genome sequencing (scWGS) data that show no evidence for somatic *APP* retrotransposition in AD neurons or in neurons from normal individuals of various ages.

We examined the original *APP*-targeted sequencing data from the Lee study to investigate sequence features of *APP* retrotransposition. These expected features included (a) reads spanning two adjacent *APP* exons without intervening intron sequence, which indicates the involvement of processed *APP* mRNA, and (b) clipped reads spanning the source *APP* and new genomic insertion sites, thus manifesting partial alignment to both the source and target site (Extended Data Fig. 1a). The first feature is the hallmark of retrogene or pseudogene insertions, and the second is the hallmark of RNA-mediated insertions of all kinds of retroelements, including retrogenes as well as LINE1 elements. We indeed observed multiple reads spanning two adjacent *APP* exons without the intron; however, we could not find any reads spanning the source *APP* and a target insertion site. Surprisingly, we found multiple clipped reads at both ends of the *APP* coding sequence (CDS) containing the multiple cloning site of the pGEM-T Easy Vector (Promega), which indicates external contamination of the sequencing library by a recombinant vector carrying an insert of *APP* coding sequence (Fig. 1a). The *APP* vector we found here was not used in the Lee study, but rather had been used in the same laboratory when first reporting genomic *APP* mosaicism[2], suggesting carryover from the prior study.

Recombinant vectors with inserts of gene coding sequences (typically without introns or untranslated regions (UTRs)) are widely used for functional gene studies. Recombinant vector contamination in next-generation sequencing is a known source of artifacts in somatic variant calling, as sequence reads from the vector insert confound those from the endogenous gene from the sample DNA[3]. We have identified multiple incidences of vector contamination in next-generation sequencing datasets from different groups, including our own laboratory (Extended Data Fig. 1b), demonstrating the risk of exposure to vector contamination. In an unrelated study on somatic copy number variation in the mouse brain[4], from the same laboratory that authored the Lee study, we found contamination of the same human *APP* pGEM-T Easy Vector in mouse single-neuron WGS data (Extended Data Fig. 1c). We also observed another vector backbone sequence (pTripIEx2, SMART cDNA Library Construction Kit, Clontech) with an *APP* insert (Extended Data Fig. 1c, magnified panel) in the same mouse genome dataset, indicating repeated contamination of multiple types of recombinant vectors in the laboratory. This highlights pervasive contamination of recombinant vector DNA in next-generation sequencing experiments, even with high quality control standards, and emphasizes the need for rigorous data analysis to mitigate this significant source of artifacts.

PCR-based experiments with primers targeting the *APP* coding sequence (e.g., Sanger sequencing and SMRT sequencing) are unable to distinguish *APP* retrocopies from vector inserts (Fig. 1a). Therefore, to definitively distinguish the three potential sources of *APP* sequencing reads (original source *APP*, retrogene copy, and vector insert), it is necessary to study non-PCR-based sequencing data (*e.g.*, SureSelect hybrid-capture sequencing) and examine reads at both ends of the *APP* coding sequence, to assess whether the clipped sequences map to a new insertion site or to vector backbone sequence. From the SureSelect hybrid-capture sequencing data in the Lee study, we directly measured the level of vector contamination by calculating the fraction of the total read depth at both ends of the *APP* coding sequence comprised by clipped reads containing vector backbone sequences (Fig. 1b, red dots). Similarly, we measured the clipped read fraction at each *APP* exon junction, which indicates the total amount of *APP* gencDNAs (either from *APP* retrocopies or vector inserts) (Fig. 1b, black dots). The average clipped read fraction at coding sequence ends (1.2%, red dots) was comparable to the average clipped read fraction at exon junctions (1.3%, black dots; *P*=0.64, Mann-Whitney U test), suggesting vector contamination as the primary source of the clipped reads across all the exon junctions. All the fractions at every junction are far below the conservative estimate of 16.5% gencDNA contribution based on their DISH experimental results (see Supplementary Information). Moreover, if the clipped reads were from endogenous retrocopies, the clipped and non-clipped reads would be expected to be of similar insert (DNA fragment) size distribution; however, we observed that in the Lee study, the clipped reads were of significantly smaller and far more homogeneous insert size distribution than the non-clipped reads that were from original source *APP*, thus demonstrating the foreign nature of the clipped reads (*P* < 2.2×10-16, Mann-Whitney U test; Extended Data Fig. 2a-c, see Supplementary Information). Finally, we found no evidence to solely support the existence of true *APP* retrogene insertions, such as clipped and discordant reads near the *APP* UTR ends that mapped to a new insertion site, or clipped reads with polyA tails at the 3' end of the UTR. All results from the hybrid-capture sequencing data suggest that the majority of *APP* gencDNA supporting reads actually originated from the *APP* vector contamination.

The Lee study reported numerous novel forms of *APP* splice variants with intra-exon junctions (IEJs) with greater diversity in AD patients than controls. The authors also presented short sequence homology (2-20 bp) at IEJs suggesting a microhomology-mediated end-joining as a mechanism underlying IEJ formation. Interestingly, IEJs were exclusively reported in the PCR-based methods, and we found no supporting evidence of any IEJs in the hybrid-capture sequencing data. It is well known that microhomology can predispose to PCR artifacts[5,6], and the Lee study performed a high number of PCR cycles in their experimental protocol (40 cycles). Thus, we tested the hypothesis that the IEJs in the Lee study could have arisen as PCR artifacts from the PCR amplification of a vector contaminant. To do so, we repeated in our laboratory both RT-PCR and PCR assays following the Lee study protocol using recombinant vectors with two different *APP* isoforms (*APP*-751, *APP*-695), and using the reported PCR primer sets with three different PCR enzymes as described in their study (see Supplementary Information). Indeed, with all combinations of *APP* inserts and PCR enzymes, we observed chimeric amplification bands with various sizes, clearly distinct from the original *APP* inserts (Fig. 1c, Extended Data Fig. 3a). We further sequenced the non-specific amplicons and confirmed that they contained numerous IEJs of *APP* inserts (Supplementary Table 1). 12 of 17 previously reported IEJs in the Lee study were also found from our sequencing of PCR artifacts (Fig. 1c and Extended Data Fig. 3b). Our observations strongly suggest that the novel *APP* variants with IEJs from the Lee study might have originated from vector contaminants as PCR artifacts.

Lastly, we examined somatic *APP* retrogene insertions in our independent scWGS data from AD patients and normal controls. Briefly, single-neuronal nuclei were isolated using FACS sorting with NeuN markers, and the extracted DNA genome was amplified with multiple displacement amplification (MDA) followed by WGS at 45X mean depth[7]. The dataset consists of a total of 64 scWGS datasets from 7 AD patients with Braak stage V and VI disease, along with 119 scWGS datasets from 15 unaffected control individuals, some of which have been previously published[8,9]. Our previous studies and those by other groups[7,10,11] have successfully detected and fully validated bona fide somatic insertions of LINE1 by capturing distinct sequence features in scWGS data, demonstrating the high resolution and accuracy of scWGS-based retrotransposition detection. Therefore, if a retrogene insertion had occurred, we should have been able to observe distinct sequence features at the source retrogene site: increased exonic read-depth, read clipping at exon junctions, poly-A tail at the end of the 3' UTR, and discordant read pairs spanning exons (Extended Data Fig. 1a). We indeed clearly captured these features at the existing germline retrogene insertions, such as the *SKA3* pseudogene insertion (Fig. 2a). If present, somatic events should be able to be detected as heterozygous germline variants in scWGS; however, our analysis revealed no evidence of somatic *APP* retrogene insertions in any of the features in any cells. We also observed a clear increase in exonic read depth relative to introns for germline retrogene insertions of *SKA3* and *ZNF100* (Fig. 2b) but observed no such read depth increase for *APP* in our 64 AD and 119 normal single-neuron WGS profiles, confirming that we found no evidence of *APP* retrogene insertions in human neurons.

In summary, our analysis of the original sequencing data from the Lee study as well as of our own scWGS data suggests that somatic *APP* retrotransposition does not frequently occur either in AD or control neurons. Rather, the reported *APP* retrocopies appear attributable to contamination by *APP* recombinant vectors. Our replication experiment also showed a possibility of PCR amplification artifacts to create spurious products that mimic *APP* gene recombination with various internal exon junctions. As noted earlier, recombinant vector contamination in next-generation sequencing is more pervasive than generally considered, warranting particularly careful data analysis. In conclusion, we found no evidence of *APP* retrotransposition in the genomic data presented in the Lee study and furthermore show that our own single-neuron WGS analysis, which directly queried the *APP* locus at single-nucleotide resolution, reveals no evidence of *APP* retrotransposition or insertion.

References

1 Lee, M. H. et al. Somatic APP gene recombination in Alzheimer's disease and normal neurons. Nature 563, 639-645, doi:10.1038/s41586-018-0718-6 (2018).

2 Bushman, D. M. et al. Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. Elife 4, doi:10.7554/eLife.05116 (2015).

3 Kim, J. et al. Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. Bioinformatics 32, 3072-3080, doi:10.1093/bioinformatics/btw383 (2016).

4 Rohrback, S. et al. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. Proc Natl Acad Sci U S A 115, 10804-10809, doi:10.1073/pnas.1812702115 (2018).

5 Odelberg, S. J., Weiss, R. B., Hata, A. & White, R. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res 23, 2049-2057, doi:10.1093/nar/23.11.2049 (1995).

6 Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLoS One 12, e0169774, doi:10.1371/journal.pone.0169774 (2017).

7 Evrony, G. D. et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron 85, 49-59, doi:10.1016/j.neuron.2014.12.028 (2015).

8 Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555-559, doi:10.1126/science.aao4426 (2018).

9 Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental

and transcriptional history. Science 350, 94-98, doi:10.1126/science.aab1785 (2015).

10 Erwin, J. A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci 19, 1583-1591, doi:10.1038/nn.4388 (2016).

11 Evrony, G. D., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. Elife 5, doi:10.7554/eLife.12966 (2016).

12 Zhang, X. et al. Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell 166, 1147-1162 e1115, doi:10.1016/j.cell.2016.07.025 (2016).

Figures

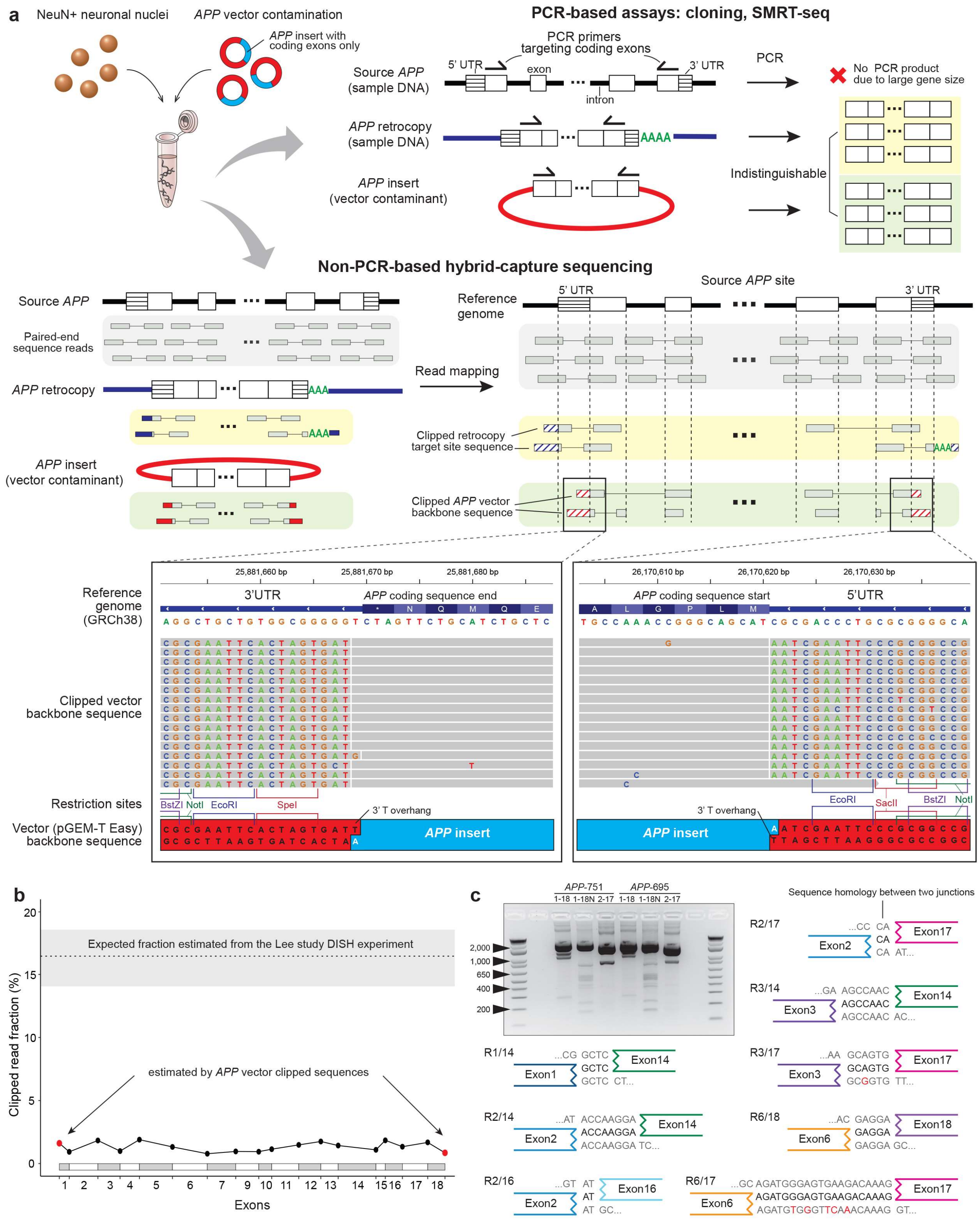

a
NeuN+ neuronal nuclei
APP vector contamination
APP insert with coding exons only
PCR-based assays: cloning, SMRT-seq
PCR primers targeting coding exons
Source APP (sample DNA)
5' UTR
exon
intron
3' UTR
PCR
No PCR product due to large gene size
APP retrocopy (sample DNA)
AAAA
APP insert (vector contaminant)
Indistinguishable
Non-PCR-based hybrid-capture sequencing
Source APP
Paired-end sequence reads
APP retrocopy
AAA
APP insert (vector contaminant)
Reference genome
Source APP site
Read mapping
Clipped retrocopy target site sequence
Clipped APP vector backbone sequence
Reference genome (GRCh38)
25,881,660 bp
25,881,670 bp
25,881,680 bp
3'UTR
APP coding sequence end
AGGCTGCTGTGGCGGGGGTCTAGTTCTGCATCTGCTC
Clipped vector backbone sequence
CGCGAATTCACTAGTGAT
Restriction sites
BstZI NotI EcoRI SpeI
3' T overhang
Vector (pGEM-T Easy) backbone sequence
CGCGAATTCACTAGTGATT
GCGCTTAAGTGATCACTAA
APP insert
26,170,610 bp
26,170,620 bp
26,170,630 bp
APP coding sequence start
5'UTR
A L G P L M
TGCCAAACCGGGCAGCATCGCGACCCTGCGCGGGGCA
AATCGAATTCCCGCGGCCG
EcoRI SacII BstZI NotI
3' T overhang
APP insert
ATCGAATTCCCGCGGCCG
TTAGCTTAAGGGCGCCGGC
b
Expected fraction estimated from the Lee study DISH experiment
estimated by APP vector clipped sequences
Clipped read fraction (%)
0
5
10
15
20
1 2 3 4 5 6 7 9 10 11 12 13 14 15 16 17 18
Exons
c
APP-751
APP-695
1-18 1-18N 2-17 1-18 1-18N 2-17
2,000
1,000
650
400
200
Sequence homology between two junctions
R2/17
...CC CA
CA
CA AT...
Exon2
Exon17
R3/14
...GA AGCCAAC
AGCCAAC
AGCCAAC AC...
Exon3
Exon14
R1/14
...CG GCTC
GCTC
GCTC CT...
Exon1
Exon14
R3/17
...AA GCAGTG
GCAGTG
GCGGTG TT...
Exon3
Exon17
R2/14
...AT ACCAAGGA
ACCAAGGA
ACCAAGGA TC...
Exon2
Exon14
R6/18
...AC GAGGA
GAGGA
GAGGA GC...
Exon6
Exon18
R2/16
...GT AT
AT
AT GC...
Exon2
Exon16
R6/17
...GC AGATGGGAGTGAAGACAAAG
AGATGGGAGTGAAGACAAAG
AGATGTGGGTTCAAACAAAG GT...
Exon6
Exon17

Figure 1. *APP* vector contamination in the Lee study. a. *APP* vector contamination and its manifestation in genome sequences. A vector *APP* insert generates indistinguishable supporting evidence of *APP* gencDNA from that generated by a true *APP* retrocopy. All designed PCR primers in the Lee study targeted only *APP* coding sequence regions shared by both *APP* retrocopy and vector *APP* insert, failing to distinguish the two sources (upper panel). In hybrid-capture sequencing, sequence reads from the flanking regions outside of the coding sequence and around the UTR regions can indicate their sources by containing the subsequence of origin (lower panel, colored in red and blue for reads originating from vectors and retrocopies, respectively). The hybrid-capture sequencing data from the Lee study clearly shows clipped reads at both ends of *APP* coding sequence with a vector backbone sequence (pGEM-T Easy), including restriction sites at the multiple cloning site, and a 3' T-overhang (magnified panel with Integrative Genomics Viewer (IGV) screenshot). The structure of the recombinant vector contaminant and its backbone sequence are depicted, showing a perfect match to the clipped sequence. PCR duplicate reads were shown together for clear visualization of read clipping. No retrotransposition-supporting reads were detected in the hybrid-capture data. **b.** Estimated fractions of cells with *APP* gencDNA at the exon junctions in the hybrid-capture data of the Lee study. All of these exon junction fractions (black dots, fractions either from retrocopies or vector inserts) are comparable to the fraction at the coding sequence ends (red dots, fractions only from the vectors), indicating that the primary source of *APP* gencDNA is vector amplification. The dotted line on the top represents the conservative estimate of expected gencDNA-supporting ratio based on the lowest occurrence rate of *APP* retrogene insertion measured in their DISH experiment (see Supplementary Methods); shaded area, 95% confidence

interval. **c.** Electrophoresis and sequencing of PCR products from the vector *APP* inserts (*APP*-751, *APP*-695) showing novel *APP* variants as artifacts. All three PCR primer sets and three PCR enzymes used in the Lee study were tested (OneStep Ahead RT-PCR, see Extended Data Fig. 3a for other results). All novel bands were further sequenced to examine the formation of IEJs with microhomology. Eight out of 12 IEJs found both in our *APP* vector PCR sequencing and RT-PCR results from the Lee study are shown (see also Extended Data Fig. 3b). Microhomology sequences are marked with reference sequences at pre- and post-junctions (grey) and sequences derived from reads (black).

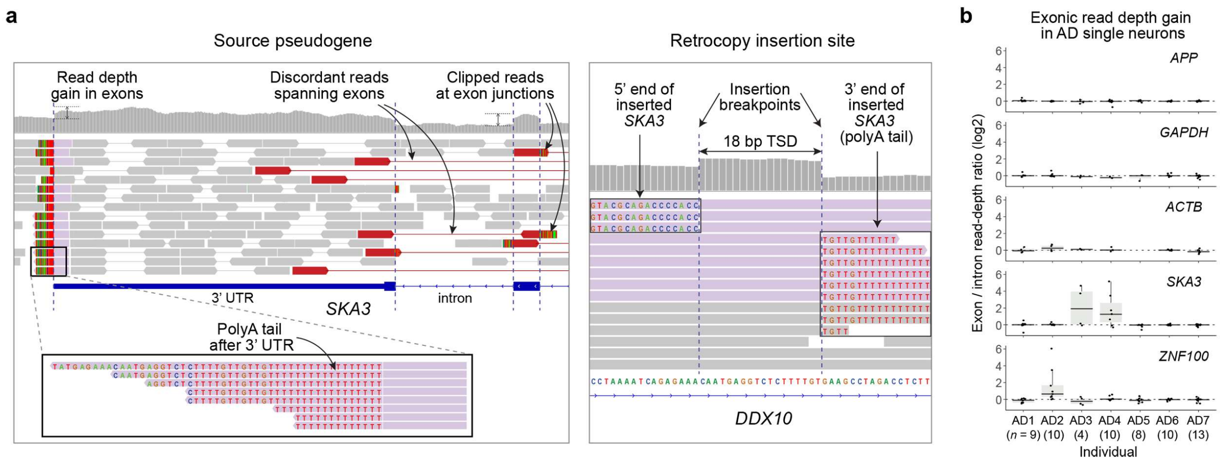
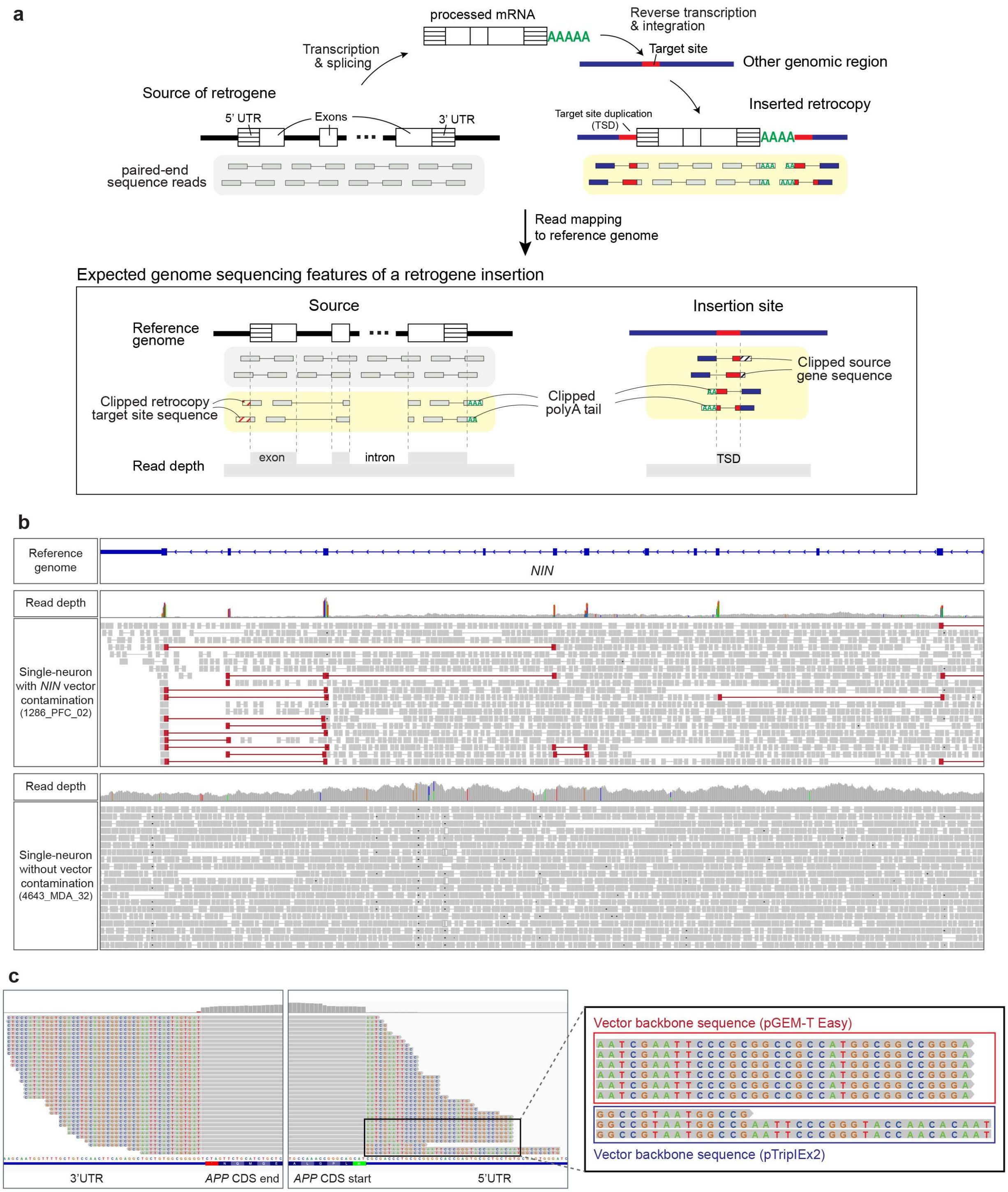
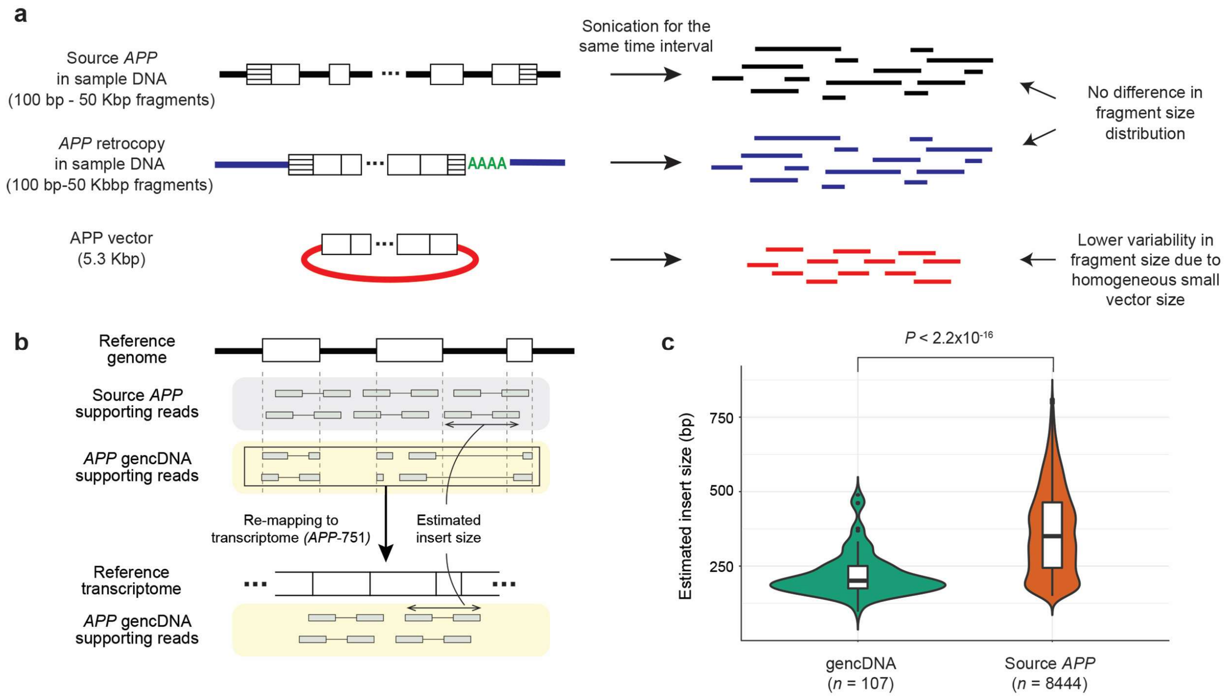


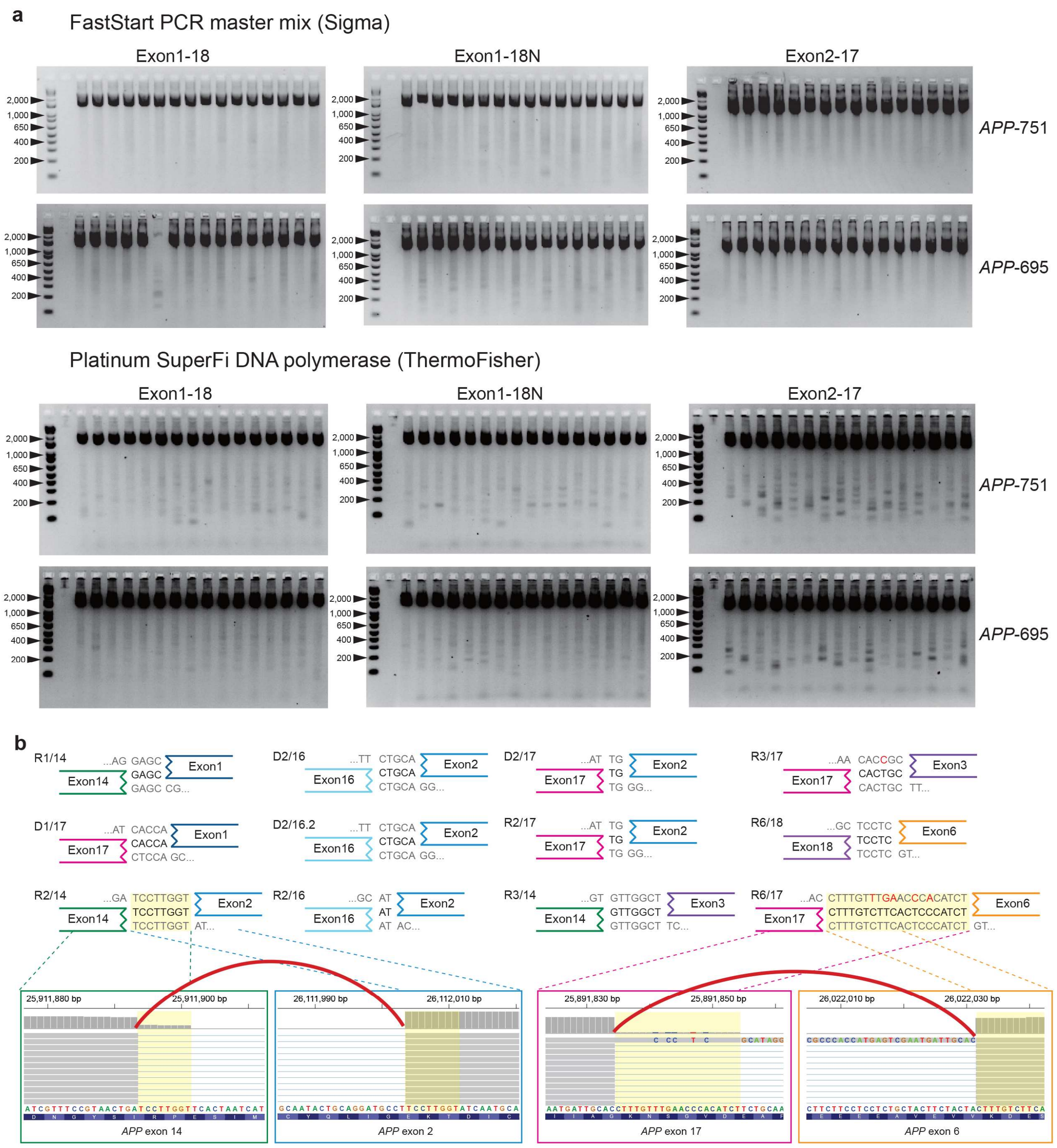
Figure 2. Absence of somatic *APP* retrogene insertions in our single-cell whole-genome sequencing data. a. An actual germline pseudogene insertion (*SKA3*) taken from our single-cell sequencing data. All distinctive characteristics including increased exonic read-depth, discordant reads spanning exons, clipped reads at exon junctions, 3' poly-A tail, and target site duplication (TSD) at the insertion site are clearly observed. Mismatches including germline single-nucleotide polymorphisms and base call errors are not shown for clear visualization of insertion characteristics. **b.** No read-depth gain in *APP* exons in our AD single neurons. Each dot represents the median of exon/intron read-depth ratios across all exons of the gene in each single neuron WGS dataset from AD patients. Along with the *APP* gene, two housekeeping genes (*GAPDH*, *ACTB*) and two source genes of germline pseudogene insertions (*SKA3* in AD3 and AD4, *ZNF100* in AD2) are depicted as negative and positive controls. Single cells that had poor genomic coverage for a given gene due to locus dropout are excluded. n, number of single cells in each individual; center line, median; box limits, first and third quartiles.



Extended Data Fig. 1. Pervasive recombinant vector contamination in next-generation sequencing. a. Schematic of a retrogene insertion and the characteristics expected to be captured in sequencing data: increased exonic read-depth, discordant reads spanning exons, clipped reads at exon junctions, 3' poly-A tail, target site duplication (TSD) at the new genomic insertion site, and clipped reads spanning the retrocopy and insertion sites. Vector contaminants can mimic most characteristics of true retrogene insertions, except for features related to new insertion sites and the insertional mechanism such as polyA tail and TSD, since recombinant vectors contain inserts of processed gene-coding sequences. **b.** Recombinant vector contamination from an experiment performed in the Walsh laboratory. Four single human neurons (1286_PFC_02, 1762_PFC_04, 5379_PFC_01, 5416_PFC_06) in our previous publication contained contamination by sequences from a mouse *Nin* recombinant vector[12]. The homologous human gene region of the source gene (*NIN*) is visualized by the IGV browser for a vector contaminated cell (upper panel) and an unaffected control cell (lower panel). Contamination characteristics including increased exonic read-depth and discordant reads spanning exons (reads colored in red) were clearly identified. Note that because the contaminant inserts were derived from the mouse *Nin* gene and mapped here on the human reference genome, numerous mismatches were observed in exonic regions (indicated by colored vertical bars in the read depth track). **c.** Another *APP* vector-contaminated dataset from the Chun laboratory[4]. This mouse single-neuron WGS data was contaminated by the same *APP* recombinant vector detected in the Lee study[1]. An additional *APP* plasmid vector was also identified in this experiment (magnified panel), suggesting contamination by multiple recombinant *APP* vectors in the laboratory.



Extended Data Fig. 2. Evidence that recombinant vector contamination is the major source of *APP* gencDNA. a. Schematic of the DNA fragment size distribution for each *APP* source (source *APP*, *APP* retrocopy, *APP* vector). Fragments from *APP* vectors are expected to be more homogeneous and smaller in size than those from other sources due to the fixed and relatively small vector size. **b.** DNA fragment (or insert) size estimation. Sequence reads mapped to *APP* exon junctions were divided into two groups: source *APP* (reads containing intron sequences) and *APP* gencDNA (reads clipped at the exon junction) supporting reads. gencDNA supporting reads were remapped to the *APP* reference transcript sequence (*APP*-751) to estimate insert sizes. **c.** Comparison of insert size distribution between source and gencDNA supporting reads. n, number of read pairs in each group.



Extended Data Fig. 3. Novel *APP* variants with intra-exon junctions as PCR artifacts. a. Electrophoresis of PCR products from the vector *APP* inserts (*APP*-751, *APP*-695) showing novel *APP* variants as artifacts. Results of two PCR enzymes (FastStart PCR master mix, Platinum SuperFi DNA polymerase) with three primer sets are presented. All combinations generated novel bands smaller than the expected PCR product. **b.** PCR-induced IEJs with homologous sequences at each junction identified by Illumina sequencing. Twelve IEJs from our vector PCR sequencing showed exactly the same sequence homologies and genomic coordinates as IEJs reported in the Lee study. For two IEJs, IGV browser images show pre- (left) and post-junction sites (right) connected by split reads spanning the IEJ (red arc). Because IGV displays forward strand sequences of the human reference genome, all IEJ sequences were also reverse complemented for consistent visualization.