Matters arising

APP gene copy number changes reflect exogenous contamination

https://doi.org/10.1038/s41586-020-2522-3

Received: 16 July 2019

Accepted: 18 May 2020

Published online: 19 August 2020

Check for updates

Junho Kim[1,2,3], Boxun Zhao[1,2,3], August Yue Huang[1,2,3], Michael B. Miller[1,2,3,4,5,6], Michael A. Lodato[1,2,3,4,5,7], Christopher A. Walsh[1,2,3,4,5] ✉ & Eunjung Alice Lee[1,2,3] ✉

ARISING FROM M. H. Lee et al. *Nature* https://doi.org/10.1038/s41586-018-0718-6 (2018)

Various types of somatic mutations occur in cells of the human body and cause human diseases, including cancer and some neurological disorders[1]. Recently, Lee et al.[2] (hereafter 'the Lee study') reported somatic copy number gains of the *APP* gene, a known risk locus for Alzheimer's disease (AD), in 69% and 25% of neurons of AD patients and controls, respectively, and argued that the mechanism of these copy number gains was somatic integration of *APP* mRNA into the genome, creating what they called genomic cDNA (gencDNA). Our reanalysis of the data from the Lee study and two additional whole-exome sequencing (WES) data sets by the authors of the Lee study[3] and Park et al.[4] revealed evidence that *APP* gencDNA originates mainly from exogenous contamination by *APP* recombinant vectors, nested PCR products, and human and mouse mRNA, respectively, rather than from true somatic integration of endogenous *APP*. We further present our own single-cell whole-genome sequencing (scWGS) data that show no evidence for somatic *APP* retrotransposition in neurons from individuals with AD or from healthy 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 would indicate processed *APP* mRNA, and (b) clipped reads, which are 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. Unexpectedly, we found multiple clipped reads at both ends of the *APP* coding sequence that contained 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[5], 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 artefacts in somatic variant calling, as sequence reads from the vector insert confound those from the endogenous gene in the sample DNA[6]. We have identified multiple incidences of vector contamination in next-generation sequencing data sets 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[7], from the same laboratory that authored the Lee study, we found contamination by 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 data set, indicating repeated contamination by multiple types of recombinant vectors in the laboratory.

PCR-based experiments with primers that target the *APP* coding sequence (for example, Sanger sequencing and SMRT sequencing) are unable to distinguish *APP* retrocopies from vector inserts (Fig. 1a, top). Therefore, to definitively distinguish between 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 (for example, SureSelect hybrid-capture sequencing) and to examine reads at both ends of the *APP* coding sequence. Such data can help to clarify whether the clipped sequences map to a new insertion site or to vector backbone sequence (Fig. 1a, bottom). 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 that consisted of 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* gencDNA (either from *APP* retrocopies or vector inserts) (Fig. 1b, black dots). The average clipped read fraction at coding sequence ends that contained vector backbones (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 that vector contamination was the primary source of the clipped reads across all the exon junctions. Even including these vector-originating reads, all the fractions at every junction are far below the conservative estimate of 16.5% gencDNA contribution based on the Lee study's DNA in situ hybridization (DISH) experimental results, which are from the same samples (see Supplementary Information for more details on the discrepancy between sequencing and DISH results). It is incumbent on the authors to provide an explanation for this inconsistency. Moreover, if the clipped reads were from endogenous retrocopies, the clipped and non-clipped reads would be expected to have a similar insert (DNA fragment) size distribution; however, in the Lee study, the clipped reads had a significantly smaller and far more homogeneous insert size distribution than the non-clipped reads that

[1]Division of Genetics and Genomics, Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA. [2]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA. [5]Department of Neurology, Harvard Medical School, Boston, MA, USA. [6]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [7]Present address: Department of Molecular, Cell, and Cancer Biology, University of Massachusetts Medical School, Worcester, MA, USA. ✉e-mail: christopher.walsh@childrens.harvard.edu; ealice.lee@childrens.harvard.edu

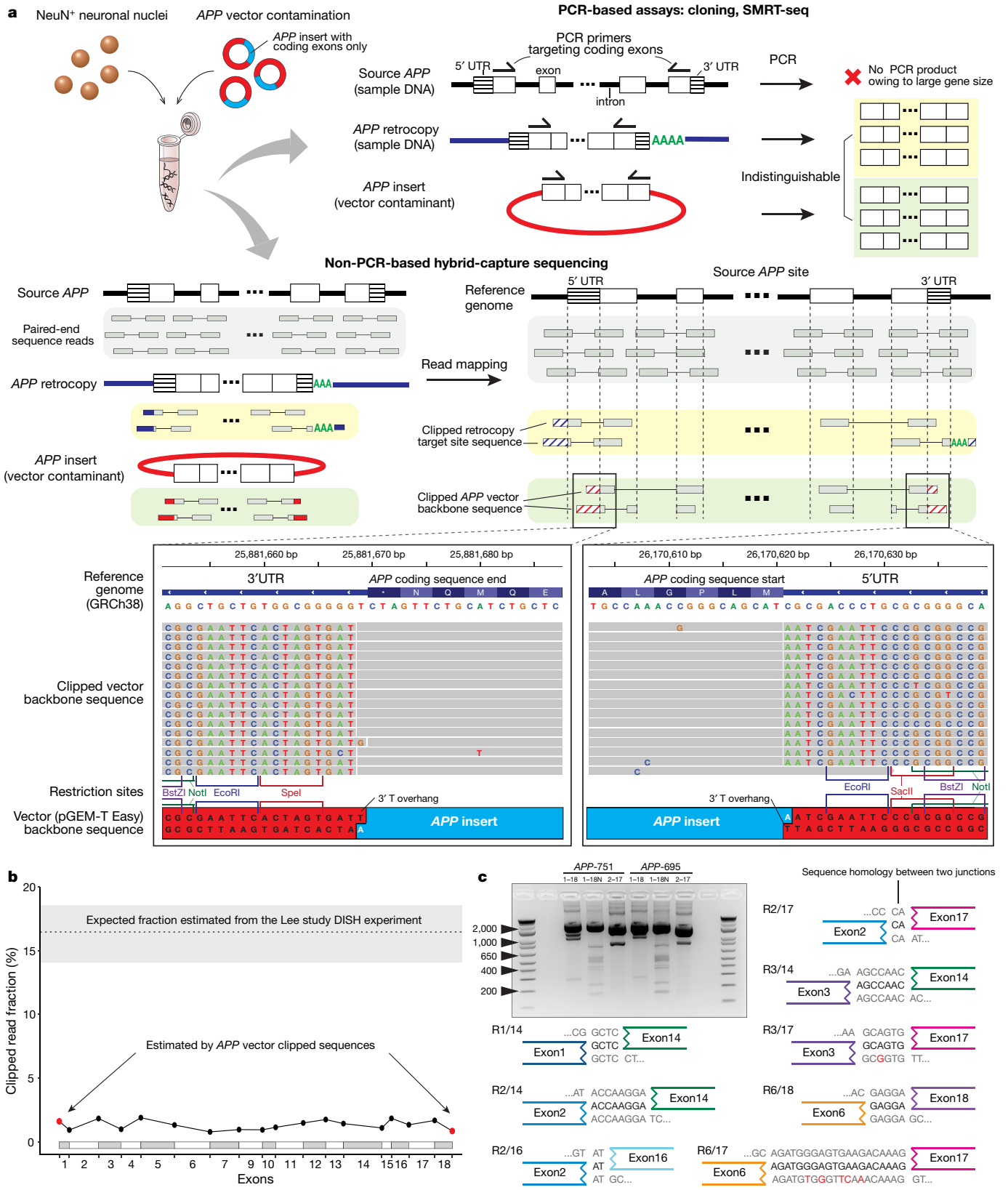


Fig. 1 | *APP* vector contamination in the Lee study. a, *APP* vector contamination and its manifestation in genome sequences. PCR-based assays in the Lee study[2] fail to distinguish between *APP* retrocopy and vector *APP* insert. Hybrid-capture sequences from the Lee study show clipped reads with a vector backbone sequence (pGEM-T Easy), including restriction sites at the multiple cloning site and a 3′ T-overhang. **b**, Estimated fractions of cells with *APP* gencDNA at the exon junctions in the Lee hybrid-capture data. All exon junction fractions (black dots) are comparable to the fraction at the coding sequence ends with vector backbone sequences (red dots). The dotted line above represents the conservative estimate of expected fraction based on the Lee DISH experiment (see Supplementary Methods); shaded area, 95% confidence interval. **c**, Electrophoresis and sequencing of PCR products from the vector *APP* inserts (*APP*-751/695) showing new *APP* variants as artefacts. Eight out of twelve IEJs found both in our *APP* vector PCR sequencing and the Lee study RT–PCR results are shown (Extended Data Fig. 3).

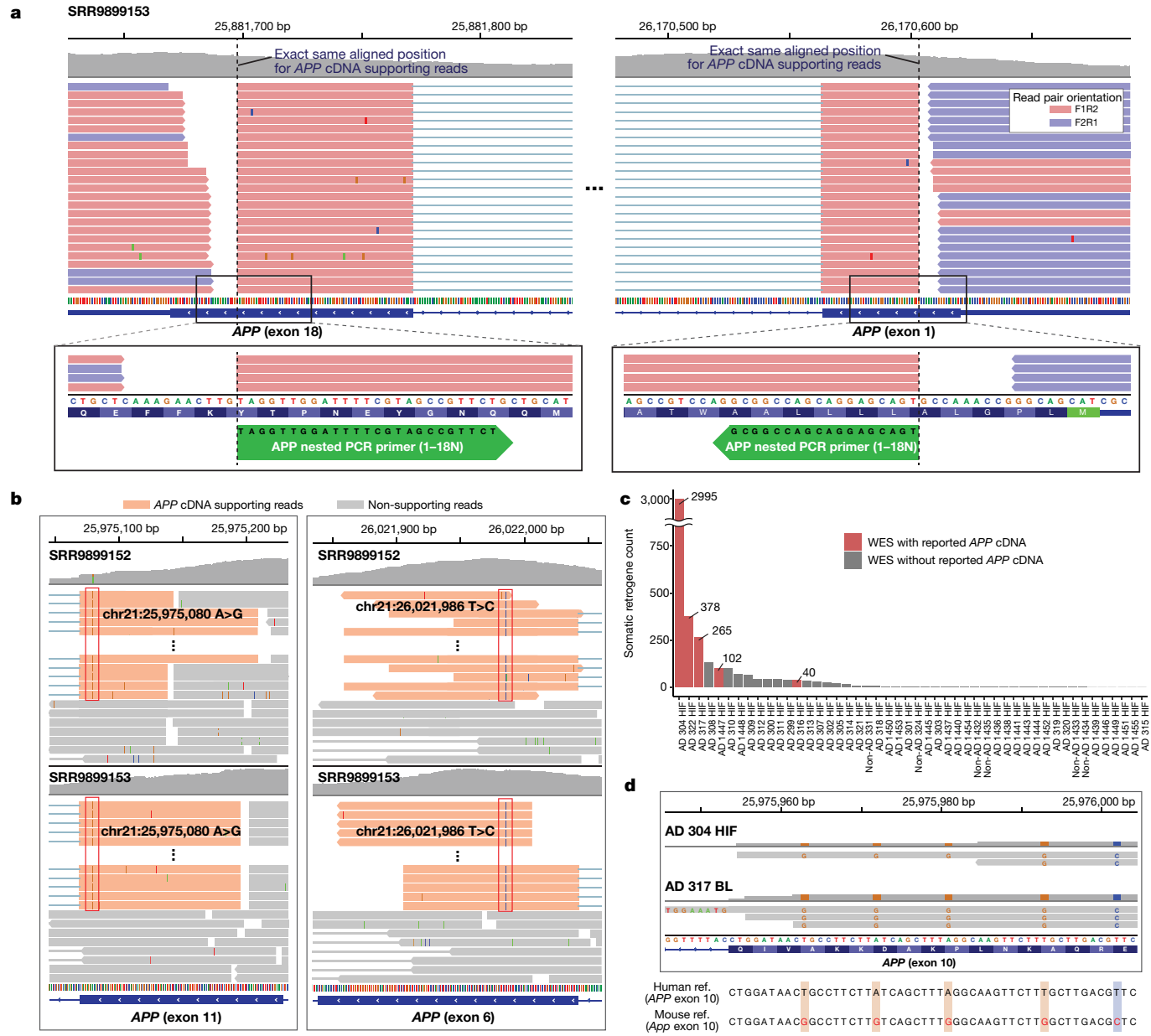


Fig. 2 | *APP* cDNA-supporting reads originate from exogenous PCR products and genome-wide human and mouse mRNA contamination. **a**, *APP* nested PCR products found in the recent Lee WES data[3]. Reads that support *APP* cDNA are aligned to the target sites (dotted lines) of the nested PCR primers (green arrows at the bottom) used in the original Lee study[2]. All these cDNA-supporting reads contain an IEJ between exons 2 and 17 (full structure not shown). **b**, The same unannotated variants found at two different positions (red boxes) only in cDNA-supporting reads (orange) in both WES data sets by Lee et al. (SRR989152 and SRR989153)[2,3]. **c**, Total gene counts with potential somatic retrogene insertions in the Park et al. data[4]. WES data with reported *APP* cDNA are marked in red. **d**, *APP* cDNA-supporting reads originating from mouse mRNA in the Park data. Mouse-specific single-nucleotide polymorphisms (coloured bases) are observed in a portion of cDNA-supporting reads, including those with clipped sequences for exon–exon junctions, suggesting the reads originated from mouse mRNA rather than genomic DNA (Supplementary Fig. 1).

were from original source *APP*, thus demonstrating the foreign nature of the clipped reads ($P < 2.2 \times 10^{-16}$, Mann–Whitney *U* test; Extended Data Fig. 2a–c, see Supplementary Information). Finally, we found no direct evidence to 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, although the sequencing depth of UTRs was over 500×. Given that the hybrid capture experiment appears properly designed to detect *APP* gencDNA, the absence of any bona fide insertion signal suggests the absence of true *APP* gencDNA and that the majority of *APP*-gencDNA-supporting reads originated from *APP* vector contamination.

The authors of the Lee study have subsequently generated WES data sets from the brain samples of six patients with AD and one control individual without AD (Sequence Read Archive (SRA) accession: PRJNA558504), and reported multiple reads spanning *APP* exons without introns as evidence of somatic *APP* gencDNA[3]. We confirmed this in the data, but again, found not a single read spanning the source *APP* and any insertion sites. Instead, the data revealed anomalous patterns in a subset of reads supporting *APP* gencDNA. Those reads spanning exons 1 and 18 were aligned to the exact same start and end positions with the same read pair orientation (Fig. 2a), which is unlikely to occur in non-PCR-based exome capture sequencing. We found that the two aligned positions within exons 1 and 18 exactly matched the target

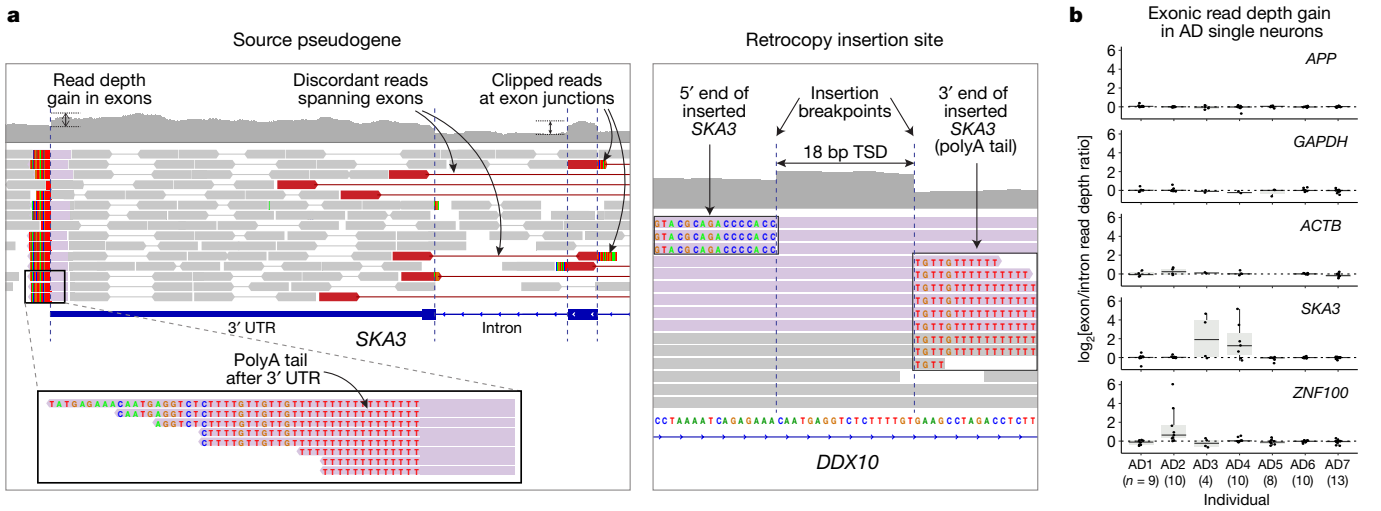


Fig. 3 | Absence of somatic *APP* retrogene insertions in our scWGS data. **a**, A germline pseudogene insertion (*SKA3*) in our scWGS data showing all distinctive characteristics of true retrogene insertion. **b**, No read-depth gain in *APP* exons in our single neurons from patients with AD. Each dot represents the median of exon/intron read-depth ratios across all exons of the gene in each scWGS data set from patients with AD. Patients with AD who have polymorphic germline retrogene insertions of *SKA3* (AD3 and AD4) or a germline insertion of *ZNF100* (AD2) show clear read-depth gain; there is no such gain for two housekeeping genes (*GAPDH*, *ACTB*). 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; centre line, median; box limits, first and third quartiles; whiskers, 1.5 × interquartile range.

sites of the nested PCR primers used in the original Lee study (1-18N, Supplementary Table 1 in the Lee study). The only explanation for this observation is contamination of the WES library by nested PCR products from the original APP study. This finding raises serious concerns that *APP* PCR products may also have contaminated the genomic DNA samples and were fragmented and sequenced together, generating more gencDNA-compatible reads for which we are unable to clarify the source. We also identified two unannotated (that is, absent in the gnomAD) single-nucleotide variants in all *APP*-cDNA-supporting reads in the two independent WES libraries pooled from six AD samples, which is very unlikely to be observed in different individuals, thus supporting the possibility that the *APP* cDNA originated from the same external source (Fig. 2b).

An independent study by Park et al.[4] has recently presented a small fraction of reads supporting *APP* cDNA in deep WES data sets from AD brain samples (SRA accession: PRJNA532465; Supplementary Fig. 12 in the study). These data were free from vector contamination, but we found evidence of genome-wide human mRNA contamination, predominantly in the WES data sets with reported *APP* cDNA supporting reads. We note that their analysis of somatic single-nucleotide variants (SNVs) is likely to be unaffected by this contamination owing to their visual inspection and stringent filtering of known germline SNVs. For each AD brain sample, we counted the number of genes with potential somatic retrotransposition events by checking whether a gene had cDNA-supporting reads (that is, reads connecting two adjacent exons and skipping the intervening intron) at more than two different exon junctions in the brain sample but not in the matched blood sample from the same patient (see Supplementary Methods). All WES data sets reported by the authors to have *APP* cDNA showed an extremely high number of other genes in addition to *APP* with cDNA-supporting reads (40–2,995 genes; Fig. 2c). Considering that far fewer than one somatic retrogene insertion per sample would be expected for human cells, even for human cancers with a high rate of somatic LINE1 retrotransposition (for example, lung and colorectal cancer)[8], this result strongly suggests that cDNA-supporting reads could not have originated from true somatic insertions of hundreds to thousands of retrogenes but rather supports the presence of genome-wide human mRNA contamination. We also found cDNA-supporting reads, including a subset of *APP* cDNA-supporting reads, that originated from mouse mRNA, additionally confirming mRNA contamination of the data (Fig. 2d, Supplementary Fig. 1). We observed mRNA contamination in one cell in our scWGS data (see Supplementary Information). Neither Park et al. (personal communication) nor we had performed any mRNA experiments, suggesting that contamination might have arisen from a source outside the research laboratories, such as the sequencing facility. We found no evidence of genuine *APP* genomic cDNA either in the new WES data from the Lee study authors, or in the independent Park et al. data. These findings highlight pervasive exogenous contamination in next-generation sequencing experiments, even with high quality-control standards, and emphasizes the need for rigorous data analysis to mitigate these important sources of artefacts.

The Lee study reported numerous new forms of *APP* splice variants with intra-exon junctions (IEJs), with greater diversity in patients with AD than in healthy individuals. The authors also presented short sequence homology (2–20 bp) at IEJs and suggested that microhomology-mediated end-joining contributed to IEJ formation. It is well known that microhomology can predispose to PCR artefacts[9], 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 artefacts from the PCR amplification of a 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 that were clearly distinct from the original *APP* inserts (Fig. 1c, Extended Data Fig. 3a). We further sequenced these non-specific amplicons and confirmed that they contained numerous IEJs of *APP* inserts (Supplementary Table 1). Twelve of seventeen previously reported IEJs in the Lee study were also found from our sequencing of PCR artefacts (Fig. 1c, Extended Data Fig. 3b). Our observations suggest that the new *APP* variants with IEJs from the Lee study might have originated from contaminants as PCR artefacts. This possibility is corroborated by the fact that IEJ-supporting reads were completely absent from the hybrid-capture sequencing data from the Lee study, and that reads supporting an IEJ in the new WES data set by the authors originated from external nested *APP* PCR products (Fig. 2a).

To independently investigate potential *APP* gencDNA, we searched for somatic *APP* retrogene insertions in our independent scWGS data from patients with AD and healthy control individuals. In brief, we isolated single neuronal nuclei using NeuN staining followed by fluorescence-activated cell sorting (FACS), amplified the whole genome using multiple displacement amplification (MDA), and finally sequenced the whole genome at 45× mean depth[10]. The dataset consists of a total of 64 scWGS data sets from 7 patients with Braak stage V and VI AD, along with 119 scWGS data sets from 15 unaffected control individuals, some of which have been previously published[11]. Our previous studies and those by other groups[10,12–14] 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 captured these features at the existing germline retrogene insertions, such as the *SKA3* pseudogene insertion (Fig. 3a). 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 cell. By contrast, in both patients (AD3 and AD4) with germline insertions of *SKA3* and the patient (AD2) with a germline insertion of *ZNF100*, there was a clear increase in exonic read depth relative to introns, as would signal for polymorphic germline retrogene insertions (Fig. 3b). We 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, the new WES data from the same authors, and the WES data from the independent Park study, as well as of our own scWGS data, suggests that somatic *APP* retrotransposition does not frequently occur in neurons from either patients with AD or healthy individuals. Rather, the reported evidence of *APP* retrocopies appears to be attributable to various types of exogenous contamination—specifically *APP* recombinant vectors, PCR products, and genome-wide mRNA contamination. Our replication experiment also showed that it is possible for PCR amplification artefacts to create spurious products that mimic *APP* gene recombination with various internal exon junctions. Thus, to support the claimed phenomenon of *APP* gencDNA, it would be necessary for the authors to present unequivocal evidence that cannot be attributed to contamination, such as reads that support new *APP* insertion breakpoints; however, the authors have not presented such direct evidence. In conclusion, we found no evidence of *APP* retrotransposition in the genomic data presented in the Lee study and further 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.

Data availability

APP vector PCR sequences have been deposited in the NCBI SRA (PRJNA577966). Single-cell whole-genome sequencing data from control individuals have been deposited in the NCBI SRA (PRJNA245456) and dbGAP (phs001485.v1.p1). Single-cell whole-genome sequencing data from patients with AD are available upon request for genomic regions of *APP* and source pseudogene *SKA3* and *ZNF100*.

Code availability

Implemented custom code for the estimation of clipped read fractions and the detection of intra-exon junctions (IEJs) is available at https://sourceforge.net/projects/somatic-app-analysis/.

1. McConnell, M. J. et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**, eaal1641 (2017).
2. Lee, M. H. et al. Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639–645 (2018).
3. Lee, M.-H. et al. Reply: *APP* gene copy number changes reflect exogenous contamination. *Nature* https://doi.org/10.1038/s41586-020-2523-2 (2020).
4. Park, J. S. et al. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat. Commun.* **10**, 3090 (2019).
5. 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**, (2015).
6. Kim, J. et al. Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics* **32**, 3072–3080 (2016).
7. 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. USA* **115**, 10804–10809 (2018).
8. Cooke, S. L. et al. Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **5**, 3644 (2014).
9. 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 (1995).
10. Evrony, G. D. et al. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
11. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
12. Erwin, J. A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci.* **19**, 1583–1591 (2016).
13. Evrony, G. D., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. *eLife* **5**, e12966 (2016).
14. Zhao, B. et al. Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genet.* **15**, e1008043 (2019).
15. Zhang, X. et al. Cell-type-specific alternative splicing governs cell fate in the developing cerebral cortex. *Cell* **166**, 1147–1162.e1115 (2016).

Acknowledgements E.A.L. is supported by grants from the NIA (K01AG051791), the Suh Kyungbae Foundation, and the Charles H. Hood foundation. This work was also supported by the Paul G. Allen Frontiers Group (C.A.W., E.A.L.), NINDS grant R01NS032457-20S1 (C.A.W.), DOD grant W18XWH2010028 (J.K., E.A.L., C.A.W.), Manton Center Pilot Project Award and Rare Disease Research Fellowship (B.Z.), NIH grants T32HL007627 and K08AG065502 (M.B.M.), and NIH grant AG054748 (M.A.L). C.A.W. is an Investigator of the Howard Hughes Medical Institute.

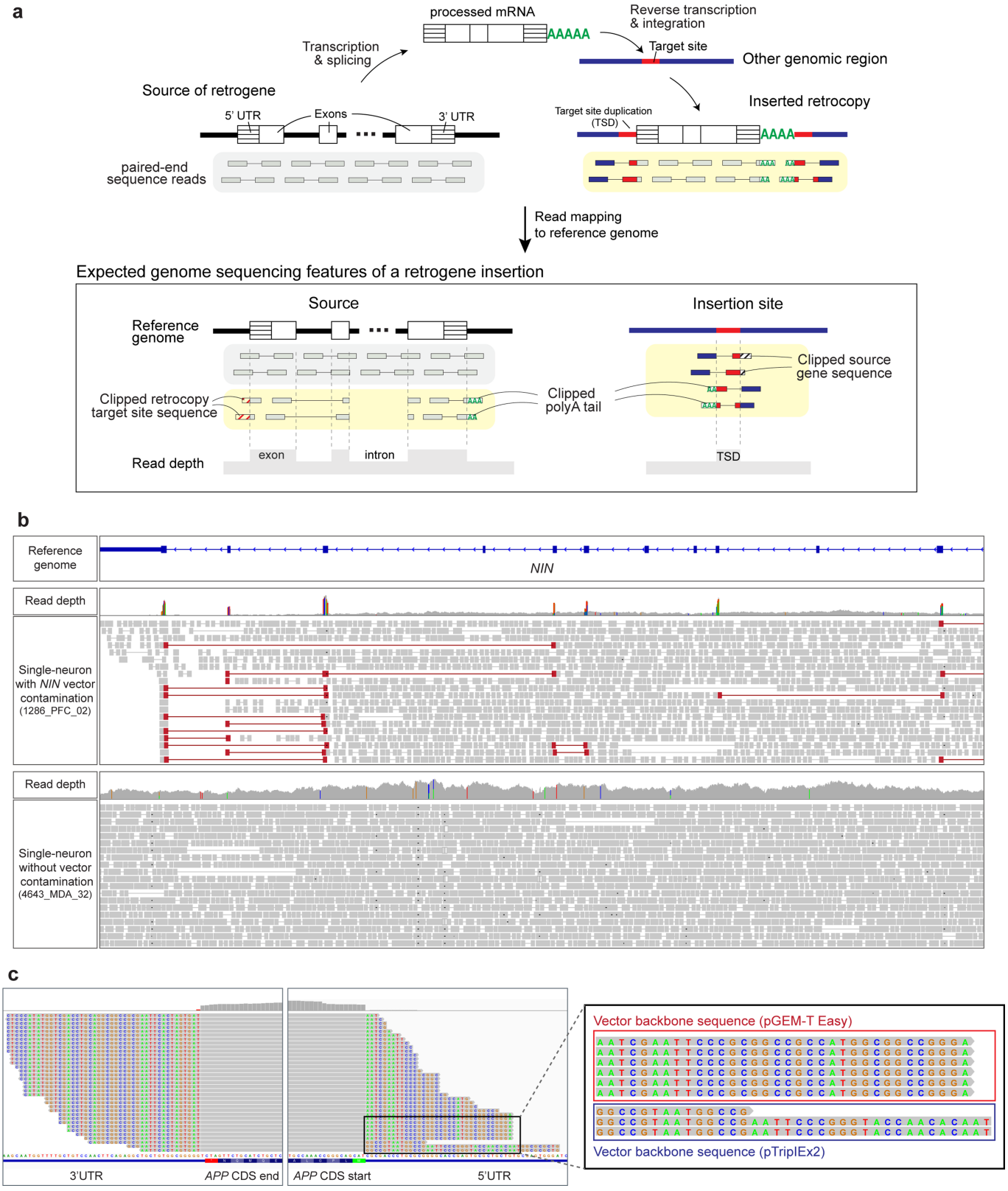
Author contributions J.K. and E.A.L. conceived and designed the study. J.K. and B.Z. designed the *APP* vector PCR and sequencing, and B.Z. performed the PCR and sequencing. M.B.M. and M.A.L. performed single-neuron sorting and sequencing. J.K. and A.Y.H. performed bioinformatic analyses. E.A.L and C.A.W supervised the study. J.K., B.Z., M.B.M., M.A.L., C.A.W., and E.A.L. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

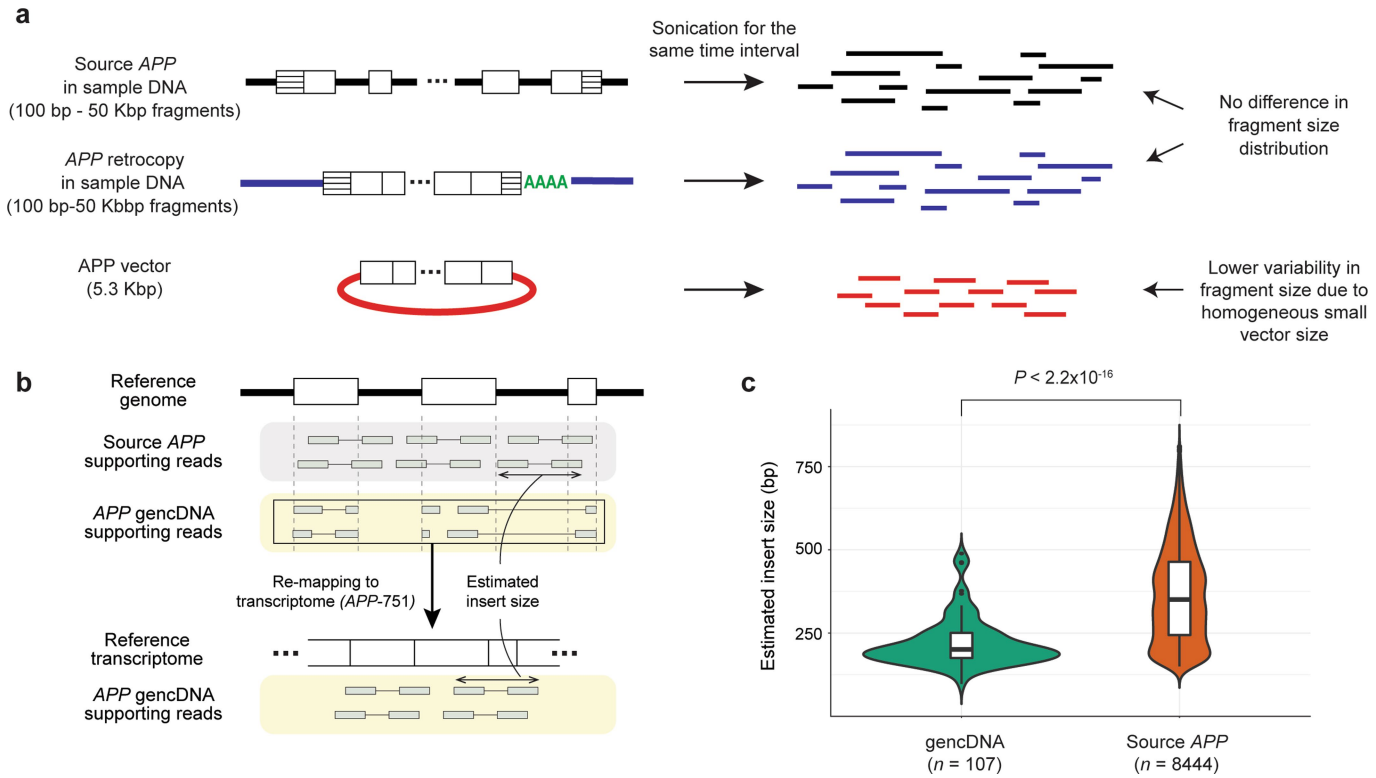
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2522-3.

Correspondence and requests for materials should be addressed to C.A.W. or E.A.L.

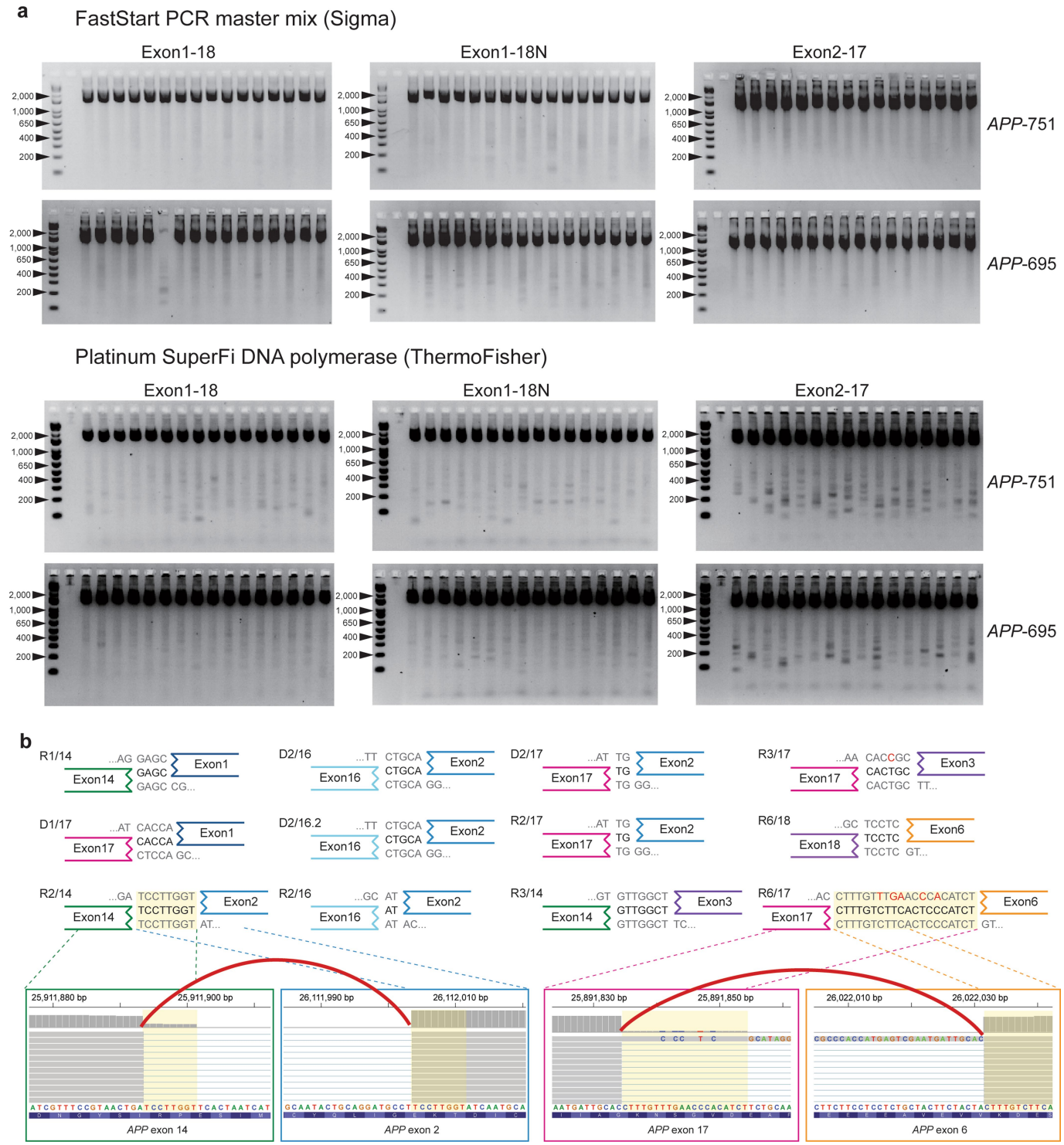


Extended Data Fig. 1 | See next page for caption.

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. **b**, Recombinant vector contamination found in the Walsh laboratory data. Four single human neurons (1286_PFC_02, 1762_PFC_04, 5379_PFC_01, 5416_PFC_06) in our previous publication showed contamination by a mouse *Nin* recombinant vector[15]. The homologous human gene region (*NIN*) is visualized by the IGV browser for a vector-contaminated cell (top) and an unaffected control cell (bottom). Contamination characteristics were identified, including increased exonic read-depth and exon-spanning discordant reads (reads coloured in red) with numerous mismatches to the human genome reference (coloured vertical bars in the read depth track). **c**, Mouse single-neuron WGS data from the Chun laboratory[7] contaminated by the same *APP* recombinant vector detected in the Lee study[2] and an additional *APP* plasmid vector (magnified panel).



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 than those from other sources owing 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; centre line, median; box limits, first and third quartiles; whiskers, 1.5 × interquartile range.



Extended Data Fig. 3 | New *APP* variants with intra-exon junctions as PCR artefacts. a, Electrophoresis of PCR products from the vector *APP* inserts (*APP*-751, *APP*-695) showing novel *APP* variants as artefacts. All combinations of two PCR enzymes (FastStart PCR master mix and Platinum SuperFi DNA polymerase; OneStep Ahead RT–PCR in Fig. 1c) with three primer sets generated new 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 by Lee et al[2]. 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.

nature research

Corresponding author(s): Eunjung Alice Lee

Last updated by author(s): Apr 27, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SRA toolkit (2.9.0) was used to download the sequencing data from the Chun laboratory (SRP162675, SRP121019) from the Sequence Read Archive as described in the Supplementary Information.
Data analysis	Sequencing data was processed to generate analysis-ready BAM using Cutadapt (1.1.4), BWA-mem (0.7.17), Picard (2.8.0), and GATK (3.5) as described in the Supplementary Information. Vecuum (1.0.1) and NCBI BLASTN were used to clarify APP vector contamination. Implemented custom code for the calculation of clipped read fractions and the detection of intra-exon junctions will be uploaded to open source repository (SourceForge).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

APP vector PCR sequences have been deposited in the NCBI Sequence Read Archive (PRJNA577966). Single-cell whole genome sequencing data of control individuals have been deposited in the NCBI Sequence Read Archive (PRJNA245456) and dbGAP (phs001485.v1.p1). Single-cell whole genome sequencing data of AD patients will be available upon request for the genomic regions of APP and source pseudogene SKA3 and ZNF100.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed our independent single-cell whole-genome sequencing (scWGS) data of AD and control neurons including previously published data sets (Lodato et al, Science, 2017). The sample size was determined by the number of sequenced cells (64 scWGS from 7 AD patients and 119 scWGS from 15 unaffected controls). This was sufficient to verify the absence of somatic APP retrotransposition, which was reported as occurring in 69% of AD neurons on average (Binomial P < 2.2e-16).
Data exclusions	One single cell (5087_MDA_02) from the public sequencing data (Lodato et al, Science, 2018) was excluded due to genome-wide mRNA contamination.
Replication	Somatic APP retrotransposition was examined in independent scWGS data from AD patients and normal controls. Both original sequencing data from the Lee study (Lee et al., Nature, 2018) and independent scWGS data show no evidence of somatic APP retrotransposition.
Randomization	Not relevant to our study since we utilized all available data sets without any allocation of samples.
Blinding	Not relevant to our study as no difference between AD and control groups was observed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Reply to: *APP* gene copy number changes reflect exogenous contamination

https://doi.org/10.1038/s41586-020-2523-2

Received: 24 April 2020

Accepted: 18 May 2020

Published online: 19 August 2020

Check for updates

Ming-Hsiang Lee[1,3], Christine S. Liu[1,2,3], Yunjiao Zhu[1], Gwendolyn E. Kaeser[1], Richard Rivera[1], William J. Romanow[1], Yasuyuki Kihara[1] & Jerold Chun[1]✉

REPLYING TO J. Kim et al. *Nature* https://doi.org/10.1038/s41586-020-2522-3 (2020)

In the accompanying comment[1], Kim et al. conclude that somatic gene recombination (SGR) and amyloid precursor protein (*APP*) genomic complementary DNAs (gencDNAs) in the brain are contamination artefacts and do not naturally exist. We disagree. Here we address the three types of analyses used by Kim et al. to reach their conclusions: informatic contaminant identification, plasmid PCR, and single-cell sequencing. Additionally, Kim et al. requested "reads supporting novel *APP* insertion breakpoints," and we now provide ten different examples that support *APP* gencDNA insertion within eight chromosomes beyond wild-type *APP* on chromosome 21 from patients with Alzheimer's disease. If SGR exists, as experimentally supported here and previously[2,3], contamination scenarios become moot.

Our informatic analyses of data generated by an independent laboratory (Park et al.)[4] complement, and are entirely consistent with, what Lee et al.[2] presented via nine distinct lines of evidence, in addition to three from a prior publication[3]. Plasmid contamination was identified in a single pull-down dataset after publication of Lee et al.[2]; however, subsequent analyses did not alter any of our conclusions, including those of our prior publications[3,5], and plasmid contamination-free replication of this approach by ourselves and others supported the original conclusions. Novel retro-insertion sites, alterations of *APP* gencDNA number and form within cell types from the same brain, and pathogenic SNVs that occur only in samples from patients with AD, all support the existence of *APP* gencDNAs produced by SGR.

One predicted outcome of SGR is the generation of novel retro-insertion sites distinct from the wild-type locus, as we demonstrated using DNA in situ hybridization (DISH; Fig. 2n in Lee et al.). Analyses of independently published data sets[4] produced by whole-exome pull-down of DNA from laser-captured human hippocampus or blood revealed ten different *APP* insertion sites within eight different chromosomes (Fig. 1, Supplementary Table 1). We identified clipped reads spanning *APP* untranslated regions (UTRs) and new genomic insertion sites on chromosomes 1, 3, 9, 10, and 12 (Fig. 1a; wild-type *APP* is located on chromosome 21). The corresponding paired-end reads mapped to the same inserted chromosome. We also identified reads spanning *APP* exon–exon junctions of gencDNAs that had mate-reads mapping to other genomic sites on chromosomes 1, 3, 5, 6, and 13 (Fig. 1b). We are unaware of contamination sources that could produce these results that are entirely consistent with our DISH data showing *APP* gencDNA locations distinct from wild-type *APP*. These new *APP* gencDNA insertion sites strongly support the natural occurrence of *APP* gencDNAs.

An *APP* plasmid contaminant (pGEM-T Easy *APP*) was found in our single pull-down dataset; however, we could not definitively determine which *APP* exon–exon reads resulted from gencDNAs as opposed to plasmid contamination, especially in view of the 11 other distinct and uncontaminated approaches that had independently supported and/or identified *APP* gencDNAs. Three other pull-down datasets from our laboratory were informatically analysed and found to contain *APP* gencDNA reads while being free from *APP* plasmid contamination by both VecScreen[6] and subsequent use of the Vecuum script[7] (Fig. 2a, b). Possible external source contamination noted by Kim et al. in two of three data sets could not definitively account for all *APP* exon–exon junctions.

The recent availability of independently generated datasets derived from patients with AD[4] provided a test for the independent reproducibility of *APP* gencDNA identification. Five brain and two blood samples from individuals with sporadic AD (SAD) contained *APP* gencDNA sequences and were shown to be plasmid-free by Vecuum[7] screening (Fig. 2a–e). In addition to exon–exon junction reads and novel insertion sites, we also identified *APP* UTR sequences paired with reads containing *APP* gencDNA exon–exon junctions (Fig. 2d, e). This may be explained by a key experimental design factor: the pull-down probes used by Park et al. contain sequences corresponding to the 5′ and 3′ UTRs of *APP*.

In addition to *APP* plasmid and amplicon contaminants, Kim et al. invoked genome-wide mouse and human mRNA contamination in the Park et al. data set. We cannot address conditions in the Park et al. laboratory but note that it is completely independent of our own. Kim et al. explain this by implicating the generation of DNA from mRNA, which requires reverse transcriptase activity. The Agilent SureSelect pull-down used by Park et al. and in our experiments do not use reverse transcriptase (Fig. 2a and Supplementary Methods), and we are unaware of any mechanism that would generate DNA from RNA in the absence of reverse transcriptase activity under the conditions used. An alternative explanation is the existence of gencDNAs that affect other genes, as we previously detected in non-*APP* intra-exonic junctions (IEJs) found in commercial cDNA Iso-Seq data sets (Extended Data Fig. 1). Additional validation would be required for new genes, but we note that an average of 450 Mb of extra DNA exists within cortical neurons from individuals with AD[3] that could accommodate new gencDNA sequences. Kim et al. invoked genome-wide mouse mRNA contamination in the Park et al. data set to account for *APP* gencDNAs, but this explanation conflicts with the available data. Mouse-specific single nucleotide polymorphisms (SNPs) in the Park et al. data set cannot account for all *APP* gencDNA-supporting reads: five of seven *APP* exon–exon junction sequences do not contain putative mouse-specific SNPs at the specific region reported by Kim et al. (Fig. 3; Kim et al. Fig. 2d). Most critically, the novel *APP* gencDNA insertion sites identified here cannot be explained by genome-wide mRNA contamination.

[1]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [2]Biomedical Sciences Program, School of Medicine, University of California San Diego, La Jolla, CA, USA. [3]These authors contributed equally: Ming-Hsiang Lee, Christine S. Liu. ✉e-mail: jchun@sbpdiscovery.org

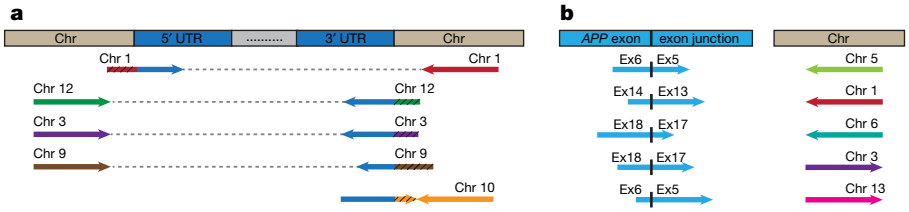


Fig. 1 | Identification of novel *APP* insertion sites in the human genome. **a**, Clipped reads spanning *APP* UTRs and novel chromosomal insertion sites were identified. The paired mate-reads of the clipped reads (black hatching) uniquely mapped to the same chromosomes. **b**, Discordant read-pairs were identified where one read spanned an *APP* exon–exon junction and the corresponding mate-read mapped to a novel chromosome. Each chromosome has a unique colour. Arrowhead direction represents the read orientation after mapping to the human reference genome. Arrows oriented in the same direction support sequence inversions. See detailed sequence and alignment information in Supplementary Table 1.

Kim et al. used PCR of *APP* splice variant plasmids, which generated sequences containing IEJs. However, there are multiple discrepancies between this approach and our biological IEJs and gencDNAs. 1) The experimental conditions, beyond the use of our primer sequences, were different: Kim et al. used twice the concentration of primers and more than one million times more template (250 pg *APP* plasmid is 4.6×10^7 copies versus about 40 gencDNA copies in our PCR of 20 nuclei; based on Lee et al.[2] Fig. 5: DISH 16/17 averaged about 1.8 copies per SAD nucleus). 2) Both gencDNA and IEJ sequences can be detected with as few as 30 cycles of PCR, as we used in single molecule real-time sequencing (SMRT-seq) (Lee et al.[2] Fig. 3) versus 40 cycles used by Kim et al. 3) The agarose gels in Kim et al. are uniformly and unambiguously dominated by a vastly over-amplified about 2-kb band (Kim et al. Fig. 1c and Extended Data Fig. 3a) that is never seen in human neurons despite our routine identification of myriad smaller bands (compare with Lee et al.[2] Fig. 2b). We did observe an over-amplified about 2-kb band in our purposeful plasmid transfection experiments, which also used PCR; however, the formation of gencDNA and IEJs was comparatively limited, of sequences distinct from brain and critically, required both reverse transcriptase activity and DNA strand breakage (Lee et al.[2], Fig. 4). 4) Finally, only 45 unique IEJs from the brains of individuals with AD and 20 from the brains of healthy controls were identified (Lee et al.[2] Fig. 3 with some overlap, fewer than 65 total) compared to the 12,426 identified by Kim et al. (an approximately 200-fold increase over biological IEJs; Kim et al. Supplementary Table 1). We wish to note that microhomology regions within *APP* exons are intrinsic to the *APP* DNA sequence and that microhomology-mediated repair mechanisms involve DNA polymerases[8,9]. The PCR results of Kim et al. differ from our biological data but might inadvertently support the endogenous formation of at least some IEJs within DNA rather than requiring RNA.

Despite these differences between the non-biological plasmid PCR data generated by Kim et al. and our data, Kim et al. conclude that IEJs from our original study[2] might have originated from contaminants. To eliminate this possibility, Lee et al.[2] presented four lines of evidence for *APP* gencDNAs containing IEJs that are independent of *APP* PCR: two different commercially produced cDNA SMRT-seq libraries, DISH, and RNA in situ hybridization (RISH). The SMRT-seq libraries revealed IEJs within *APP* (Lee et al.[2] Extended Data Fig. 1e) as well as other genes (Extended Data Fig. 1), which cannot be attributed to plasmid contamination or PCR amplification. The DISH and RISH results support the existence of *APP* gencDNAs and IEJs (see Supplementary Discussion and Lee et al.[2] Fig. 2, Extended Data Figs. 1, 2) by using custom-designed and validated commercial probe technology (Advanced Cell Diagnostics, ACD), which was independently shown to detect exon–exon junctions[10] and single-nucleotide mutations[11]. Thus, gencDNAs and IEJs can be detected in the absence of targeted PCR. Notably, the contamination proposed by Kim et al. cannot account for the marked change in the number and forms of *APP* gencDNAs that occurs with disease state. The change is also apparent when comparing cell types; signals are vastly

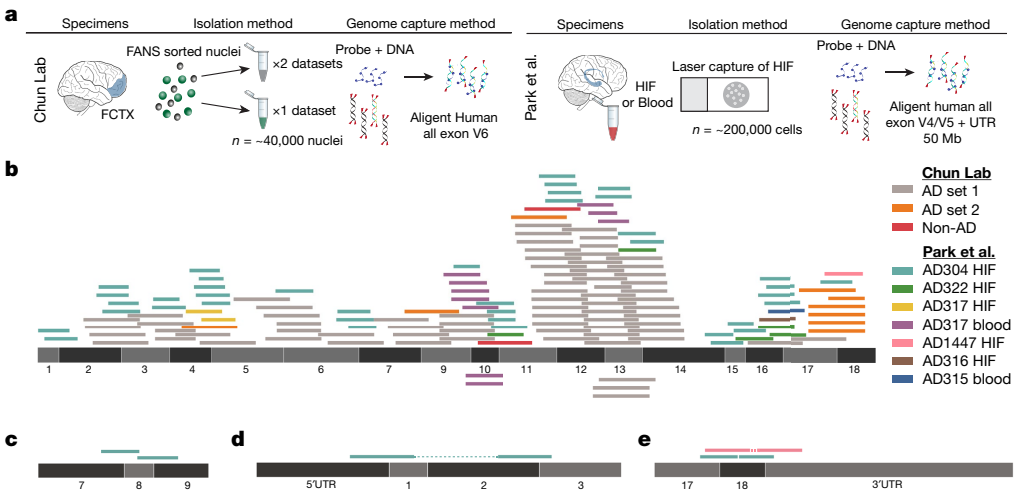


Fig. 2 | Identification of *APP* gencDNA sequences in ten new whole-exome pull-down datasets from two independent laboratories. **a**, Method schematic depicting the standard protocol for whole-exome pull-downs and highlighted methodological differences between the independent laboratories (our lab and Park et al.[4]). **b**, *APP-751* sequence with non-duplicate gencDNA reads from the ten new datasets; colour key indicates the source reads for all panels. **c**, Reads that map to junctions between *APP* exons 7, 8, and 9 that are absent from *APP-751*. **d**, **e**, Paired reads that represent a DNA fragment containing both an exon–exon junction and an *APP* 5′ or 3′ UTR.

Fig. 3 | Five *APP* gencDNA-supporting reads that span exon–exon junctions and do not contain mouse-specific SNPs. *APP* gencDNA reads were identified that span the *APP* exon10–exon11 junction from the Park et al. datasets[4]. The reference sequences of human and mouse exons are indicated and the positions where the nucleotides differ are highlighted. Five of the seven exon–exon junction-spanning reads do not contain mouse-specific SNPs.

more prevalent in neurons than in non-neuronal cells from the same brains of individuals with SAD when the samples are processed at the same time by DISH (Lee et al.[2] Fig. 5). Independent peptide nucleic acid fluorescence in situ hybridization (PNA-FISH) and dual-point-paint experiments from our previous work further support *APP* gencDNAs[3] (Table 1). Critically, SMRT-seq identified 11 single-nucleotide variations that are considered pathogenic in familial AD and that were present only in our samples from individuals with SAD; none of them exist as plasmids in our laboratory.

Kim et al. compared *APP* gencDNA copy number estimates from pull-down sequencing and DISH. However, a direct comparison is not possible since the two methodologies are fundamentally different. For example, pull-downs use solution hybridization on isolated DNA, whereas DISH uses solid-phase hybridization on fixed and sorted single nuclei. Moreover, the sequences targeted are not the same. Pull-down probes target wild-type sequences for endogenous and gencDNA loci, resulting in pull-down competition. By contrast, DISH probes target only gencDNA sequences to provide greater sensitivity. Competition by

Table 1 | Summary of targeted and non-targeted *APP* PCR methods and lines of evidence that support *APP* gencDNAs and IEJs

	Method	Targeted *APP* PCR	Support for the existence of IEJs and gencDNAs	Reference
Approaches without targeted *APP* PCR				
1	RISH on IEJ 3/16	None	IEJ 3/16 RNA signal is present in human SAD brain tissue	Lee et al.[2]
2	Whole-transcriptome SMRT-seq	None	An independent commercial source identified IEJs in *APP* and other genes	Public dataset[a], Lee et al.[2] this Reply
3	Targeted RNA SMRT-seq	None	RNA pull-down that identified *APP* IEJs	Public dataset[a], Lee et al.[2]
4	DISH of gencDNAs	None	IEJ 3/16 and exon–exon junction 16/17 showed increases in neurons compared to non-neurons from the same brain from an individual with SAD and to non-diseased neurons; J20 mice containing the *APP* transgene under a PDGF-β-promoter showed increased number and size of signal compared to non-neurons and wild-type mice	Lee et al.[2]
5	Dual point-paint FISH	None	Identified *APP* CNVs of variable puncta size that were not always associated with Chr21	Bushman et al.[3]
6	PNA-FISH	None	*APP* exon copy number increases show variable signal size and shape with semiquantitative exonic probes	Bushman et al.[3]
7	Agilent SureSelect targeted pull-down	None	Identified *APP* gencDNAs in brains from individuals with SAD; contains plasmid sequence contamination	Lee et al.[2], this Reply
New #7	Agilent all-exon pull-down	None	All-exon pull-downs, with no plasmid contamination by both Vecscreen and Vecuum, contain *APP* gencDNA sequences and evidence of gencDNA UTRs and novel insertion sites	Park et al.[4], this Reply
Approaches with targeted *APP* PCR				
8	RT–PCR and Sanger sequencing	Oligo-dT primed and targeted *APP* primers	Novel *APP* RNA variants with IEJs; predominantly in neurons from individuals with SAD	Lee et al.[2]
9	Genomic DNA PCR and Sanger sequencing	Yes	Identified *APP* gencDNAs with IEJs; predominantly in neurons from individuals with SAD	Lee et al.[2]
10	Genomic DNA PCR and SMRT-seq	Yes	IEJ/gencDNAs were more prevalent in number and form in neurons from individuals with SAD compared to non-diseased neurons; identified 11 pathogenic SNVs that were present only in SAD samples	Lee et al.[2]
11	APP-751 overexpression in CHO cells	Yes	IEJ and gencDNA formation required DNA strand breakage and reverse transcriptase	Lee et al.[2]
12	Single-cell qPCR	Yes; individual exon	Intragenic exon 14 single-cell qPCR showed copy number increases in prefrontal cortical neurons over cerebellar neurons from the same brain of an individual with SAD	Bushman et al.[3]

CNV, copy number variation.
[a]The Alzheimer brain Iso-Seq dataset was generated by Pacific Biosciences, Menlo Park, California. Additional sequencing information and analysis is provided at https://downloads.pacbcloud.com/public/dataset/Alzheimer_IsoSeq_2016/.

wild-type loci reduces the efficiency of capture, which is underscored by 32% to 40% of nuclei that do not contain gencDNAs and would contribute only wild-type sequences (Lee et al., Fig. 5c, f'). Moreover, a majority of gencDNA positive nuclei (62% to 73%) showed two or fewer signals (Lee et al., Fig. 5c, f') which reduced the relative representation of gencDNA loci. As IEJs do not contain the full exon sequence, there is inefficient hybridization and a lack of sequence capture and detection. This limitation is overcome by SMRT-seq (Table 1). Lastly, multiple other protocol variations exist, including tissue preparation, fixation, and hybridization conditions, which explain the hypothesized discrepancies.

Kim et al.'s third type of analysis yielded a negative result via interrogation of their own single-cell whole-genome sequencing (scWGS) data, which cannot disprove the existence of *APP* gencDNAs. An average of nine neurons from the brains of seven individuals with SAD were examined, raising immediate sampling issues required to detect mosaic *APP* gencDNAs. Kim et al. self-identified "uneven genome amplification"[1,12–14] that resulted in about 20% of their single-cell genomes having less than 10× depth of coverage[14] with potential amplification failure at one (~9% allelic dropout rate) or both alleles (~2.3% locus dropout rate)[12,14]. These limitations are compounded by potential amplification biases reflected by whole-genome amplification failure rates that may miss neuronal subtypes and/or disease states, which is especially relevant to single copies of *APP* gencDNAs that are as small as about 0.15 kb (but still detectable by DISH). Kim et al. state that the increased exonic read depth relative to introns reliably detects germline retrogene insertions in single cells from affected individuals (Kim et al., Fig. 3b); however, these data also demonstrate that increased exonic read depth is not observed in all cells—or even a majority in some cases—from the same individuals carrying the germline insertions of *SKA3* (AD3 and AD4) and *ZNF100* (AD2). These results demonstrate inherent technical limitations in the work by Kim et al. that prevent the accurate detection of even germline pseudogenes present in all cells, thus explaining an inability to detect the rarer mosaic gencDNAs produced by SGR. Kim et al.'s informatic analysis is also based on the unproven assumption that the structural features of gencDNA are shared with processed pseudogenes and LINE1 elements (Kim et al. Fig. 3a and Extended Data Fig. 1a), and possible differences could prevent straightforward detection under even ideal conditions as has been documented for LINE1[15]. These issues could explain Kim et al.'s negative results.

Considering these points, we believe that our data and conclusions supporting SGR and *APP* gencDNAs remain intact and warrant their continued study in the normal and diseased brain.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data from Park et al. were deposited in the National Center for Biotechnology Information Sequence Read Archive database under accession number PRJNA532465. Data from the newly reported full exome pull-down data sets will be provided for the *APP* locus upon request.

Code availability

The source codes of the customized algorithms are available on GitHub at https://github.com/christine-liu/exonjunction.

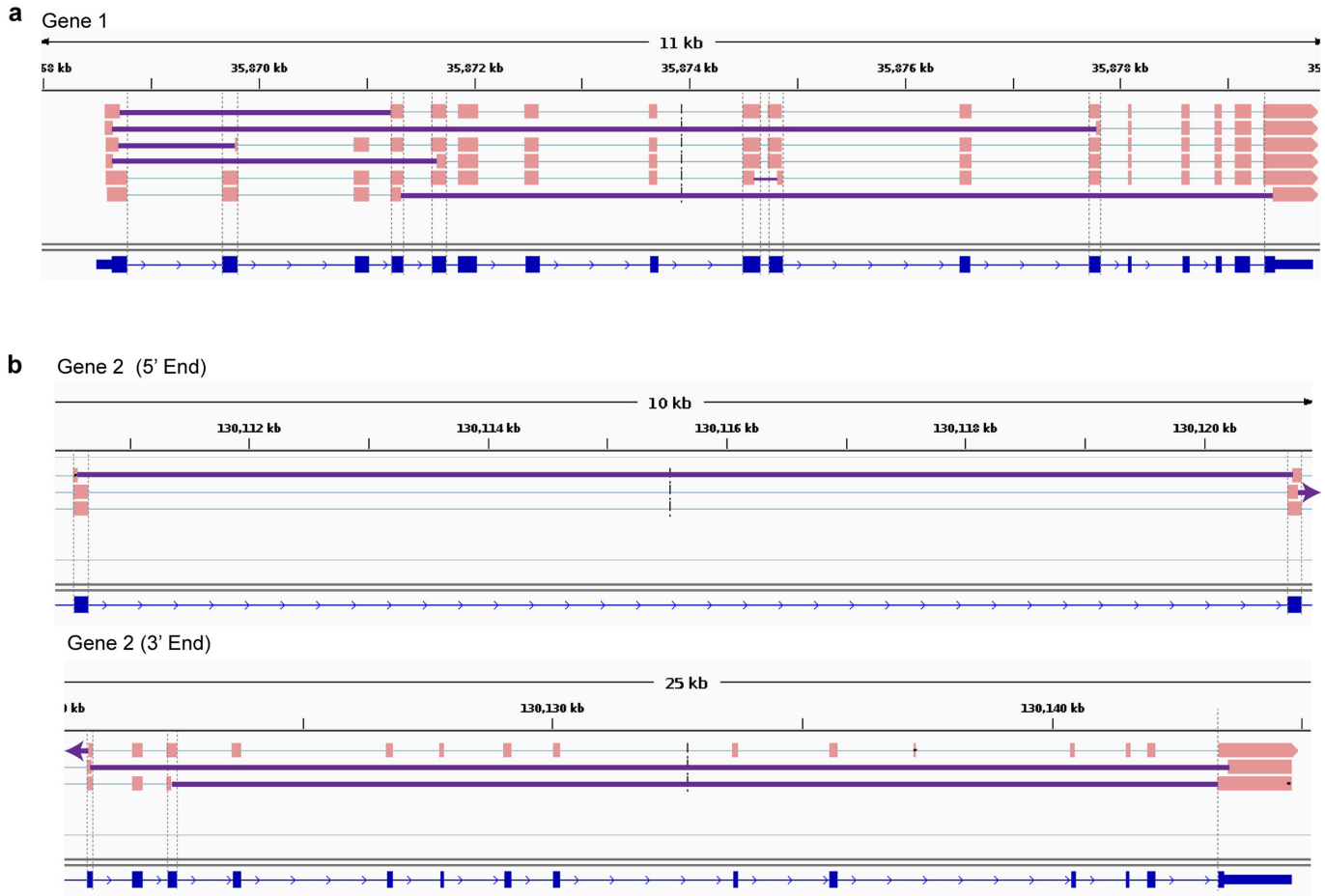
1. Kim, J. et al. APP gene copy number changes reflect exogenous contamination. *Nature* https://doi.org/10.1038/s41586-020-2522-3 (2020).
2. Lee, M. H. et al. Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639–645 (2018).
3. Bushman, D. M. et al. Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *eLife* **4**, e05116 (2015).
4. Park, J. S. et al. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat. Commun.* **10**, 3090 (2019).
5. 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. USA* **115**, 10804–10809 (2018).
6. Cummings, J. L., Morstorf, T. & Zhong, K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res. Ther.* **6**, 37 (2014).
7. Kim, J. et al. Vecuum: identification and filtration of false somatic variants caused by recombinant vector contamination. *Bioinformatics* **32**, 3072–3080 (2016).
8. van Schendel, R., van Heteren, J., Welten, R. & Tijsterman, M. Genomic scars generated by polymerase theta reveal the versatile mechanism of alternative end-joining. *PLoS Genet.* **12**, e1006368 (2016).
9. Sfeir, A. & Symington, L. S. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem. Sci.* **40**, 701–714 (2015).
10. Splice variant case study: EGFRvIII detection in glioblastoma. https://acdbio.com/science/applications/research-areas/egfrviii (ACD, 2019).
11. Baker, A. M. et al. Robust RNA-based in situ mutation detection delineates colorectal cancer subclonal evolution. *Nat. Commun.* **8**, 1998 (2017).
12. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
13. Cai, X. et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014).
14. Evrony, G. D. et al. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
15. Rohrback, S., Siddoway, B., Liu, C. S. & Chun, J. Genomic mosaicism in the developing and adult brain. *Dev. Neurobiol.* **78**, 1026–1048 (2018).

Acknowledgements We thank L. Wolszon and D. Jones for manuscript editing. Research reported in this publication was supported by the NIA of the National Institutes of Health under award numbers R56AG067489 and P50AG005131 (J.C.) and NINDS R01NS103940 (Y.K.). This work was supported by non-federal funds from The Shaffer Family Foundation, The Bruce Ford & Anne Smith Bundy Foundation, and Sanford Burnham Prebys Medical Discovery Institute funds (J.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions M.-H.L., Y.K., W.J.R. and R.R. conducted laboratory experiments; C.S.L. and Y.Z. analysed sequencing data; and J.C. conceived and oversaw the experiments. G.E.K, C.S.L, and Y.Z. created figures. All authors wrote and edited the manuscript. This Reply was the work of current laboratory members.

Competing interests Sanford Burnham Prebys Medical Discovery Institute has filed the following patent applications on the subject matter of this publication: (1) PCT application number PCT/US2018/030520 entitled, 'Methods of diagnosing and treating Alzheimer's disease' filed 1 May 2018, which claims priority to US provisional application 62/500,270 filed 2 May 2017; and (2) US provisional application number 62/687,428 entitled, 'Anti-retroviral therapies and reverse transcriptase inhibitors for treatment of Alzheimer's disease' filed 20 June 2018. J.C. is a co-founder of Mosaic Pharmaceuticals.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2523-2.
Correspondence and requests for materials should be addressed to J.C.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | IEJs identified from commercially available long-read transcriptome datasets in two genes other than *APP*. Sequences containing IEJs were identified and shown for gene 1 (**a**) and gene 2 (**b**). Gene 2 is shown in two parts. Grey dashed lines show ends of RefSeq exons; solid purple lines denote IEJs. All splice isoforms were examined. The Alzheimer brain Iso-Seq dataset was generated by Pacific Biosciences, Menlo Park, CA, and additional information about the sequencing and analysis is available at https://downloads.pacbcloud.com/public/dataset/Alzheimer_IsoSeq_2016/.

nature research

Corresponding author(s): Jerold Chun

Last updated by author(s): February, 21 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina sequencing of AD/MS datasets: Illumina NextSeq 500. Fastq files for Park et al. datasets were downloaded from SRA (accession PRJNA532465).
Data analysis	Sequences were aligned to the human reference genome (GRCh38) using STAR (version 2.5.3a) with the settings: --outSAMattributes All --outSJfilterCountTotalMin 1 1 1 1. Duplicate reads were marked and removed using Picard (version 2.1.1). Reads were then processed and visualized using a modified version of the R exonjunction package (https://github.com/christine-liu/exonjunction). Datasets were also analyzed using Vecuum (version 1.0.1) to confirm that APP plasmid was not detected in all of these datasets.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Fastq files of the Illumina short read sequences used in the analysis will be provided upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes indicated in figures and text were determined based on the availability of post-mortem human brain samples and the experience of the authors.
Data exclusions	No data was excluded from analysis.
Replication	All attempts at replication were successful.
Randomization	Samples were allocated randomly.
Blinding	No blinding procedure has been applied.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All antibodies used are listed (clone number, dilution, supplier, catalog number) Rabbit monoclonal anti-NeuN antibody (27-4, 1:800, Millipore, MABN140) Alexa Fluor 488 donkey anti-rabbit IgG antibody (N/A, 1:500, Invitrogen, Ref# A21206)
Validation	These antibodies are all published and validated by immunofluorescence staining (anti-NeuN, anti-rabbit), immunohistochemistry (anti-NeuN), and Western blot (anti-NeuN). Additional validation and peer-reviewed papers are available on the manufacturer's websites.