Abrupt warming events in the Arctic pull the meteorological equator — a band of tropical storm clouds that circle the globe near the Equator — farther north, and, along with it, the rainfall patterns associated with the Asian summer monsoon. In 2017, a reinterpretation of water-isotope signals in an Antarctic ice core identified a near-instantaneous response of atmospheric circulation to changes in Arctic climate that occurred in the most recent ice age, all the way south to West Antarctica[8]. However, whether this response occurred throughout the Southern Hemisphere, or was more localized, remained unclear.

Buizert and colleagues present the first Antarctic-wide evidence for a rapid atmospheric coupling of the position of the westerly winds around the whole of the Southern Ocean to past abrupt climate events in the Arctic (Fig. 1). Identifying these pervasive fluctuations in wind position, which happened on a decadal timescale tens of thousands of years ago, required the precise synchronization of ages for ice cores from across the Antarctic continent.

Ice-core ages from Greenland have been linked to those from Antarctica using the methane composition of bubbles in the exceedingly well-resolved ice core from the West Antarctic Ice Sheet Divide[5]. Atmospheric methane is quickly mixed across the hemispheres, and so can be considered as globally synchronous. Past fluctuations in methane abundance mimicked abrupt changes in Greenland temperature and therefore provide a way of precisely interrogating the timing of climate events between the Arctic and the Antarctic.

Buizert *et al.* took the next step in synchronizing the West Antarctic Ice Sheet Divide record with four other Antarctic ice cores by identifying characteristic sequences of volcanic eruptions preserved in the sulfate levels in Antarctic ice. Only then were the authors able to identify the superimposed oceanic and atmospheric signals that occurred across Antarctica in response to past rapid changes in Arctic climate.

The classic see-saw of heat between the hemispheres through the ocean can explain the delayed and gradual changes in Antarctic temperature that accompanied past abrupt shifts in Greenland temperature. But Buizert and co-workers' study suggests that superimposed on these slow ocean changes were an almost synchronous northward shift in the westerly winds circling Antarctica when Greenland moved into its warm phase — and, vice versa, a southward shift in these winds during cool Greenland events. This atmospheric response modulated the latitude in the Southern Ocean that formed the source of the moisture that fell as snow over Antarctica.

A one-to-one relationship has previously been identified between the duration of Greenland temperature events and the magnitude of the ensuing temperature response in Antarctica through the ocean mechanism[7]. Similarly, the authors find that the atmospheric response seems to scale so that stronger Greenland events result in a larger climatic signal in Antarctica and the Southern Ocean. An atmospheric link tying changes in Arctic climate to the Antarctic has previously been hypothesized on the basis of climate-model responses in experiments designed to mimic aspects of Dansgaard–Oeschger events[6]. The current work provides the observational data to prove the existence of this link.

It is time to move beyond considering only the Atlantic Ocean and century-scale time lags when thinking about how the Arctic and the Antarctic are climatically connected[5]. Buizert and colleagues' identification of a rapid atmospheric link between climates at the poles has implications for our understanding of current climate change. Today, the Arctic is warming at about twice the rate of the global average; however, continent-scale warming of the Antarctic that is expected from climate simulations has not yet been clearly observed[9,10]. Changes in Antarctic sea ice are also not following expectations based on models[10]. Meanwhile, the westerly winds of the Southern Hemisphere have been shifting rapidly southwards, affecting water security in cities such as Perth in Australia and Cape Town in South Africa, and potentially having global consequences by altering the movement of heat and carbon dioxide between the atmosphere and the ocean[11].

Many challenges remain in accurately predicting how, and how quickly, the behaviour of Antarctica and the Southern Ocean will change in a warming climate. Nevertheless, the authors have provided a glimpse of the natural changes in behaviour — both rapid and slow — that occurred tens of thousands of years ago. These results provide a basis for progress in unravelling the current scientific mysteries of how the ocean and the atmosphere at the poles respond to rapid changes in climate. ■

Nerilie J. Abram *is at the Research School of Earth Sciences and the Centre of Excellence for Climate Extremes, Australian National University, Canberra, ACT 2601, Australia.*
e-mail: nerilie.abram@anu.edu.au

1. Fischer, H. *et al. Nature Geosci.* **11**, 474–485 (2018).
2. Kindler, P. *et al. Clim. Past* **10**, 887–902 (2014).
3. Buizert, C. *et al. Nature* **563**, 681–685 (2018).
4. Stocker, T. F. & Johnsen, S. J. *Paleoceanography* **18**, 1087 (2003).
5. WAIS Divide Project Members. *Nature* **520**, 661–665 (2015).
6. Pedro, J. B. *et al. Quat. Sci. Rev.* **192**, 27–46 (2018).
7. EPICA Community Members. *Nature* **444**, 195–198 (2006).
8. Markle, B. R. *et al. Nature Geosci.* **10**, 36–40 (2017).
9. Abram, N. J. *et al. Nature* **536**, 411–418 (2016).
10. Jones, J. M. *et al. Nature Clim. Change* **6**, 917–926 (2016).
11. Toggweiler, J. R. & Russell, J. *Nature* **451**, 286–288 (2008).

ALZHEIMER'S DISEASE

A mosaic mutation mechanism in the brain

Variable brain-specific mutations have been observed in Alzheimer's disease. One mechanism underlying this mosaicism involves integration of variant gene copies back into the neuronal genome. SEE ARTICLE P.639

GUOLIANG CHAI & JOSEPH G. GLEESON

Genetic mutations can arise not only in fertilized eggs, affecting all cells of an organism, but also in a subset of an organism's cells[1–3]. The latter phenomenon, called mosaicism, is prevalent in the brain, and has been associated with several neurological disorders, including sporadic Alzheimer's disease, the most common form of the disease[1,3,4]. In 2015, it was found[5] that neurons from people with sporadic Alzheimer's contained more DNA and had more copies of the Alzheimer-related gene *amyloid-β precursor protein* (*APP*) than did neurons from people without the disease. However, the exact genomic changes underlying this mosaicism remained unresolved. Lee *et al.*[6] follow up on that work on page 639, providing a mechanism for increased *APP* mosaicism in the brains of people with sporadic Alzheimer's disease. The study could alter our understanding of the roots of neurodegeneration.

First, Lee *et al.* set out to analyse *APP* variants in neuronal messenger RNA. In each experiment, the authors used mRNA from just 50 neurons from the brains of people with or without sporadic Alzheimer's, because averaging across large neuronal populations could mask variants present in only a few cells. The researchers' analysis revealed many *APP* mRNA variants. As expected, the variants lacked introns — non-protein-coding regions that are removed during gene transcription through a process called splicing, leaving only protein-coding exons. However, the variants were shorter than expected, and contained single-nucleotide mutations, inserted and

deleted exons, and larger deletions that led to the formation of new exon–exon junctions between missing multi-exon regions. Some of the mutations the authors observed have been previously implicated in familial Alzheimer's disease[7].

Lee and colleagues found the same short variants when they analysed genomic DNA from the neurons, suggesting that *APP*-variant mRNAs might be transcribed from matching genomic DNA sequences — named genomic complementary DNAs (gencDNAs) by the authors — that had become permanently embedded in the genomes of neurons. To further validate the existence of *APP* gencDNAs in neurons, the authors used two independent approaches: a technique called DNA *in situ* hybridization (DISH), in which fluorescent molecules were bound to gencDNA-specific exon–exon junctions in DNA; and sequencing of short sections of *APP* DNA. Both approaches confirmed the existence of gencDNA variants.

The researchers next investigated the extent of gencDNA diversity using DNA sequencing. In total, they identified 6,299 different *APP* gencDNA variants in 96,424 neurons from the brains of 5 people with sporadic Alzheimer's — approximately 10 times more than they found in the brains of people without the disease. In agreement, DISH also revealed substantially more gencDNAs in Alzheimer's neurons.

The authors demonstrated that *APP* gencDNAs were present in the neurons of a mouse model of Alzheimer's disease, but rarely in non-neuronal cells or neurons from control animals. Moreover, gencDNA variants accumulated with age. These findings are consistent with a role for *APP* gencDNA variants in the development of Alzheimer's. Indeed, the authors found that some *APP* mRNA variants are translated into proteins that are toxic to cells, further strengthening this possibility.

Finally, Lee and co-workers showed that gencDNAs could be generated in cells in culture, provided that two conditions were met. First, the cells' DNA had to contain breaks in its strands, and, second, the enzyme reverse transcriptase had to be active. This enzyme is responsible for a process called reverse transcription, in which matching DNA sequences are produced from mRNA. The data indicate that gencDNAs arise from reverse-transcribed mRNA intermediates, which are incorporated into the genome in a process that might be promoted by breaks in DNA (Fig. 1). In support of this idea, the authors detected reverse transcriptase activity in the human brain samples, and a previous study has shown the presence of DNA breaks in developing brains[8], whereas this phenomenon is rarely observed in other tissue types.

The incorporation of gencDNAs into the genome might share some mechanisms with retrotransposition — a process in which RNA transcribed from DNA sequences called transposable elements can reintegrate into new genomic regions to generate mosaicism[3]. But how gencDNAs become mutated from the original *APP* sequence remains unknown. Perhaps the mutations arise from mis-splicing of mRNA, or during genomic integration of gencDNAs.

Taken together, Lee and colleagues' work reveals the surprising existence of a phenomenon known as somatic gene recombination in the brain. This phenomenon, which has previously been reported only in antibody generation in immune cells[9], increases the diversity of proteins encoded by a given gene through DNA-shuffling mechanisms.

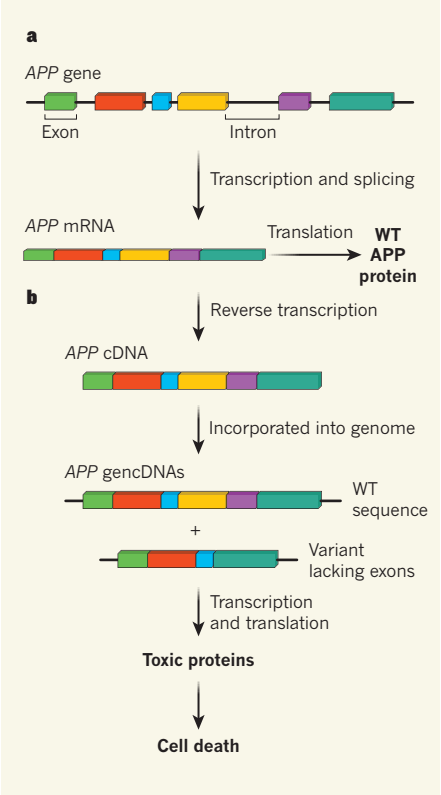


Figure 1 | Mosaic incorporation of *APP* variants into the neuronal genome. **a**, The gene *amyloid-β precursor protein* (*APP*) contains protein-coding exons (coloured blocks) and non-coding introns (this simplified schematic of the gene does not reflect the actual exon–intron composition). During transcription, introns are removed through a process called splicing to produce messenger RNA, which is translated to form the wild-type (WT) protein. **b**, Lee and colleagues[6] found that, in neurons in the human brain, *APP* mRNA undergoes a process called reverse transcription to produce a complementary DNA (cDNA). The cDNA can be reintegrated into the neuronal genome as a genomic cDNA (gencDNA). At some point in the process, mutations arise — perhaps when the cDNA is integrated into the genome, or at an earlier stage (not shown). This results in a range of gencDNA *APP* variants, some lacking one or more exons. Some gencDNA variants give rise to toxic proteins, leading to cell death. These processes might contribute to sporadic Alzheimer's disease.

The study hints at a previously unanticipated mechanism in the development of Alzheimer's, and expands our understanding of the genesis of brain mosaicism. But whether accumulation of gencDNAs in neurons is a cause of or is caused by Alzheimer's disease remains to be proved.

The techniques used here could be applied to investigate whether gencDNA mechanisms are at work in other genes in other tissues; this could provide insights into diseases such as cancer or other degenerative disorders. However, it remains possible that gencDNA production is specific to *APP* or to neurons. The authors did not find gencDNA variants in another gene involved in Alzheimer's, *presenilin*, but nor did they rule out the possibility that gencDNAs could arise from other genes. Neurons have many features that might make them particularly vulnerable to gencDNAs: they are long-lived, have mostly stopped dividing, and have higher levels of reverse transcriptase activity and DNA-strand breaks than do non-neuronal cells[8].

It is also unclear whether the integration of *APP* gencDNAs into DNA is random or is biased towards certain genomic regions. The development of more-powerful sequencing techniques should help to answer this question.

Of course, there are many other avenues for further research. For instance, whether gencDNAs co-opt the retrotransposition and integration pathways used by transposable elements remains to be tested. The fact that gencDNAs are found in normal neurons suggests that they could have some benefits — this possibility should be examined. Finally, it will be interesting to test whether inhibitors of reverse transcriptase can prevent the accumulation of gencDNAs. Only when these avenues have been explored will we be able to build a complete picture of the remarkable phenomenon observed by Lee and colleagues. ■

Guoliang Chai *and* **Joseph G. Gleeson** *are in the Department of Neurosciences, Howard Hughes Medical Institute, University of California, San Diego, La Jolla, California 92093, USA, and at the Rady Children's Institute for Genomic Medicine, Rady Children's Hospital, San Diego.*
e-mail: jogleeson@ucsd.edu; gchai@ucsd.edu

1. Biesecker, L. G. & Spinner, N. B. *Nature Rev. Genet.* **14**, 307–320 (2013).
2. Freed, D., Stevens, E. L. & Pevsner, J. *Genes* **5**, 1064–1094 (2014).
3. Richardson, S. R., Morell, S. & Faulkner, G. J. *Annu. Rev. Genet.* **48**, 1–27 (2014).
4. Beck, J. A. *et al. Hum. Mol. Genet.* **13**, 1219–1224 2004).
5. Bushman, D. M. *et al. eLife* **4**, e05116 (2015).
6. Lee, M.-H. *et al. Nature* **563**, 639–645 (2018).
7. Murrell, J., Farlow, M., Ghetti, B. & Benson, M. D. *Science* **254,** 97–99 (1991).
8. Blaschke, A. J., Staley, K. & Chun, J. *Development* **122**, 1165–1174 (1996).
9. Tonegawa, S. *Nature* **302**, 575–581 (1983).

This article was published online on 21 November 2018.

ARTICLE

https://doi.org/10.1038/s41586-018-0718-6

Somatic *APP* gene recombination in Alzheimer's disease and normal neurons

Ming-Hsiang Lee[1], Benjamin Siddoway[1,3], Gwendolyn E. Kaeser[1,2,3], Igor Segota[1,3], Richard Rivera[1], William J. Romanow[1], Christine S. Liu[1,2], Chris Park[1,2], Grace Kennedy[1], Tao Long[1] & Jerold Chun[1]*

The diversity and complexity of the human brain are widely assumed to be encoded within a constant genome. Somatic gene recombination, which changes germline DNA sequences to increase molecular diversity, could theoretically alter this code but has not been documented in the brain, to our knowledge. Here we describe recombination of the Alzheimer's disease–related gene *APP*, which encodes amyloid precursor protein, in human neurons, occurring mosaically as thousands of variant 'genomic cDNAs' (gencDNAs). gencDNAs lacked introns and ranged from full-length cDNA copies of expressed, brain-specific RNA splice variants to myriad smaller forms that contained intra-exonic junctions, insertions, deletions, and/or single nucleotide variations. DNA in situ hybridization identified gencDNAs within single neurons that were distinct from wild-type loci and absent from non-neuronal cells. Mechanistic studies supported neuronal 'retro-insertion' of RNA to produce gencDNAs; this process involved transcription, DNA breaks, reverse transcriptase activity, and age. Neurons from individuals with sporadic Alzheimer's disease showed increased gencDNA diversity, including eleven mutations known to be associated with familial Alzheimer's disease that were absent from healthy neurons. Neuronal gene recombination may allow 'recording' of neural activity for selective 'playback' of preferred gene variants whose expression bypasses splicing; this has implications for cellular diversity, learning and memory, plasticity, and diseases of the human brain.

The diversity of neuronal form and function is intrinsic to the human brain, but its basis remains largely unknown. Early speculations involved gene recombination[1], analogous to the mechanism of antibody diversification that was later identified[2], but this has not been described in the brain[3,4]. Nevertheless, later identification of genomic mosaicism[5], which arises somatically to produce brain cells with distinct if seemingly random genomic changes, suggested genome dynamism that might include gene recombination. Genomic mosaicism was first identified in neural progenitor cells and neurons as aneuploidies and DNA content variation, both representing large copy number variations (CNVs)[6–8]. Randomly distributed, smaller megabase-scale CNVs, LINE1 repeat elements, and single nucleotide variations (SNVs) were subsequently identified. Genomic mosaicism can influence cell survival and gene transcription, but somatic gene recombination of specific genes has not been reported[5,9].

A candidate gene for neuronal recombination is *APP*, which shows mosaic CNVs in normal human brains. These CNVs are increased in sporadic Alzheimer's disease (SAD)[10], the most common form of Alzheimer's disease. *APP* is central to the amyloid hypothesis wherein APP is cleaved by secretases to form toxic amyloid-β (Aβ) peptides and plaques, causing Alzheimer's disease[11]. Constitutive *APP* mutations and duplications are believed to cause rare forms of familial Alzheimer's disease (FAD) and Alzheimer's disease neuropathology in Down syndrome (trisomy 21 with 3 *APP* copies), supporting the idea that they have a pathogenic role when present mosaically in SAD[12–14]. We previously identified mosaic, neuronal *APP* CNVs that showed heterogeneous signals that might be explained by gene recombination[10]. However, interrogation of *APP* genomic loci (about 0.3 Mb) using low-depth, short-read single-cell sequencing capable of detecting CNVs produced negative results that were complicated by resolution limitations[5,15]. We therefore developed an alternative strategy focused on *APP* in small cell populations, using nine distinct methodologies (Extended Data Table 1).

Novel *APP* RNA variants in neurons

We postulated that genomic sequence alterations in *APP*, existing mosaically, could be detected in RNA through transcriptional amplification. Assessments were focused on small populations of nuclei rather than bulk samples that are dominated by annotated species (Extended Data Fig. 1a) to detect mosaic alterations. The workflow (Fig. 1a) commenced with fluorescence-activated nuclear sorting (FANS)[16] to isolate neuronal nuclei from prefrontal cerebral cortices from both control individuals and those with verified SAD, which were run in parallel (Extended Data Table 2). Groups of 50 NeuN-positive neuronal nuclei were isolated and processed for PCR with reverse transcription (RT–PCR; Fig. 1a) and downstream analysis. RT–PCR using validated primers on exon 1 and exon 18 (Supplementary Table 1), which can amplify full-length *APP* cDNA (*APP*-770, NM_000484.3), detected the expected splice variants *APP*-751 (NM_201413.2) and *APP*-695 (NM_201414.2)[17] (Extended Data Fig. 1b). However, multiple unexpected bands of varied sizes were also identified (Fig. 1b). The RT–PCR products were Southern blotted with ^{32}P-labelled APP cDNA probes (Fig. 1c), and positive bands were cloned and Sanger sequenced. The new bands yielded *APP* cDNA sequence variants unlike any previously reported, characterized by loss of central exons with proximal and distal exons linked by intra-exonic junctions (IEJs) (Fig. 1d,e). Twelve novel RNA variant sequences with unique IEJs were identified in neurons (Fig. 1e) and non-neurons displayed no variants (Extended Data Fig. 1c). IEJs were independently observed in five oligo-dT-primed cDNA libraries; three from sorted neuronal nuclei from individuals with SAD (Extended Data Fig. 1d) and two from commercially

[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: Benjamin Siddoway, Gwendolyn E. Kaeser, Igor Segota. *e-mail: jchun@sbpdiscovery.org

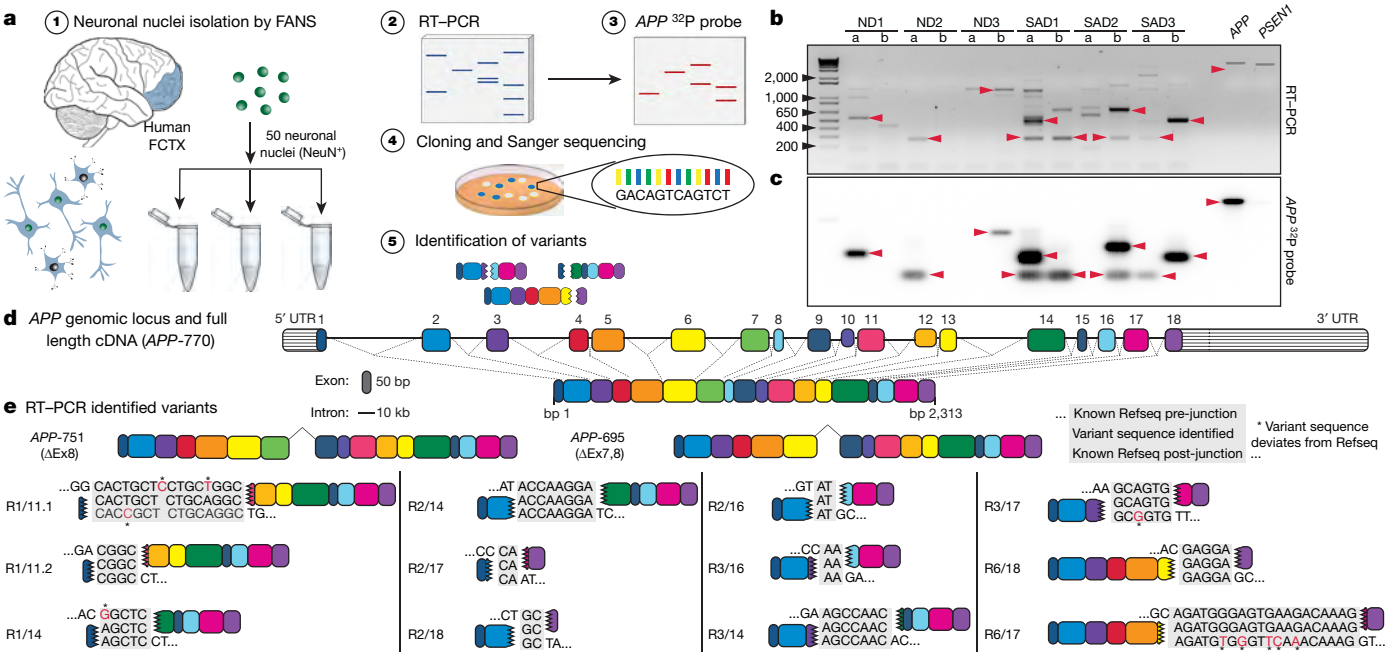


Fig. 1 | Identification of novel APP RNA variants from small populations of neurons. a, Fifty neuronal nuclei were sorted from human prefrontal cortices (FCTX) (1) and used for RT–PCR (2). The resulting RT–PCR products were screened by Southern blotting with ^{32}P-labelled *APP* cDNA probes (3). Bands with positive signals from duplicate gels were cloned and sequenced (4), and variants were identified (5). **b**, Electrophoresis of RT–PCR products from the brains of three non-diseased (ND) individuals and three patients with SAD, with two populations each (a and b). *APP* and *PSEN1* plasmids were run as positive and negative controls for Southern blotting. **c**, Southern blot of RT–PCR products. Arrowheads indicate examples of bands from **b** that were cloned and Sanger sequenced. **d**, Structure of human *APP* genomic locus and spliced *APP*-770 full-length cDNA drawn to scale; the colour scheme remains consistent throughout all figures. **e**, *APP* RNA variants identified by RT–PCR. The sequences of homology regions forming IEJs are shown. Variant sequences deviating from Refseq are shown in red with asterisks. R, RNA identified; #/#, exon–exon junction; .#, for multiple unique junctions.

produced long-read RNA-seq data sets from whole brain and temporal lobe of patients with SAD (Extended Data Fig. 1e). Five variants retained coding potential and seven contained premature stop codons (Extended Data Table 3, Supplementary Table 2). One prevalent form was characterized by an IEJ between the 24th nucleotide of exon 3 and the 45th nucleotide of exon 16 (Fig. 1e, R3/16). Detection of R3/16 by RNA in situ hybridization (RISH) on SAD brain sections indicated the cytoplasmic presence of variants (Extended Data Fig. 1f). Notably, sequence complementarity of joined exons was found in all 12 IEJs, ranging in overlap from 2 to 20 nucleotides (Fig. 1e, Extended Data Table 3, Supplementary Table 2). Amplification for a second gene related to Alzheimer's disease, *PSEN1*, did not identify variants (Extended Data Fig. 1g).

gencDNA sequences in neuronal genomes

The existence of previously unidentified RNA variants raised the question of whether this transcriptional heterogeneity originated from mosaic variation in DNA. We carried out high-stringency amplification, using the *APP* primers previously used for RNA and cDNA analyses, on RNase-treated DNA extracted from sets of 20 neuronal nuclei from both healthy brains and those with SAD (Fig. 2a). PCR of the wild-type *APP* genomic locus was not possible because of its length (about 300 kb) (Fig. 1d). However, PCR on genomic DNA generated similar-sized bands to novel RNA variants (Fig. 2b, Extended Data Fig. 2a). Sanger sequencing revealed multiple gencDNAs and seven of eight were identical to those identified in RNA (Fig. 2c). We validated the presence of *APP* gencDNAs in neurons using multiple, distinct primer sets (Extended Data Fig. 2b, c). We did not detect gencDNAs in DNA isolated from human lung fibroblasts (IMR-90), human embryonic kidney cells (HEK-293), or non-neuronal nuclei from the brains of individuals with or without SAD (Extended Data Fig. 2d, e). Amplification of *PSEN1* did not produce products from genomic DNA (Fig. 2b, Extended Data Fig. 2a).

gencDNA detection by non-PCR methods

To validate the presence of *APP* gencDNA junctions within single neuronal genomes without polymerase-based amplification, we developed DNA in situ hybridization (DISH). Our method extensively modified the sample preparation and hybridization protocols (see Methods) of a commercial RISH product, BaseScope (ACD), to recognize genomic sequences. BaseScope technology uses paired ISH probes to eliminate hybridization artefacts and can detect specific junctions. Two DISH probes were extensively used (Extended Data Table 4): one that recognized a common gencDNA sequence via the exon 16–exon 17 junction ($DISH_{16/17}$), which spans the Aβ coding region of *APP*; and one that recognized the newly identified IEJ formed between exons 3 and 16 ($DISH_{3/16}$). Bound probes were visualized as red dots with varying diameters. All probes passed multiple specificity requirements involving positive and negative controls. Sense and antisense DISH probes produced similar results in RNase-treated neuronal nuclei from individuals with SAD (Fig. 2d–i). By comparison, RNA signals were detected only using the antisense probes (Extended Data Fig. 1f); therefore, sense probes were used in all subsequent DISH analyses. Critically, DISH signals were eliminated by destruction of the target sequence by specific (but not off-target) restriction enzyme digestion (Fig. 2j–m, Extended Data Fig. 2f). In addition, no DISH signal was detected on cells infected with retroviruses containing wild-type human genomic *APP* sequences lacking target sequences (Extended Data Fig. 2g, h). Notably, double labelling with dual DISH probes recognizing the intron 2–exon 3 wild-type genomic sequence combined with $DISH_{3/16}$ or $DISH_{16/17}$ demonstrated that *APP* gencDNAs did not usually co-localize with the wild-type locus (Fig. 2n). Thus, DISH detected specific *APP* gencDNA junctions within genomic DNA without polymerase-dependent amplification, revealing multiple loci distinct from germline *APP* alleles.

A completely independent approach also identified *APP* gencDNAs without primary PCR amplification by using a custom Agilent

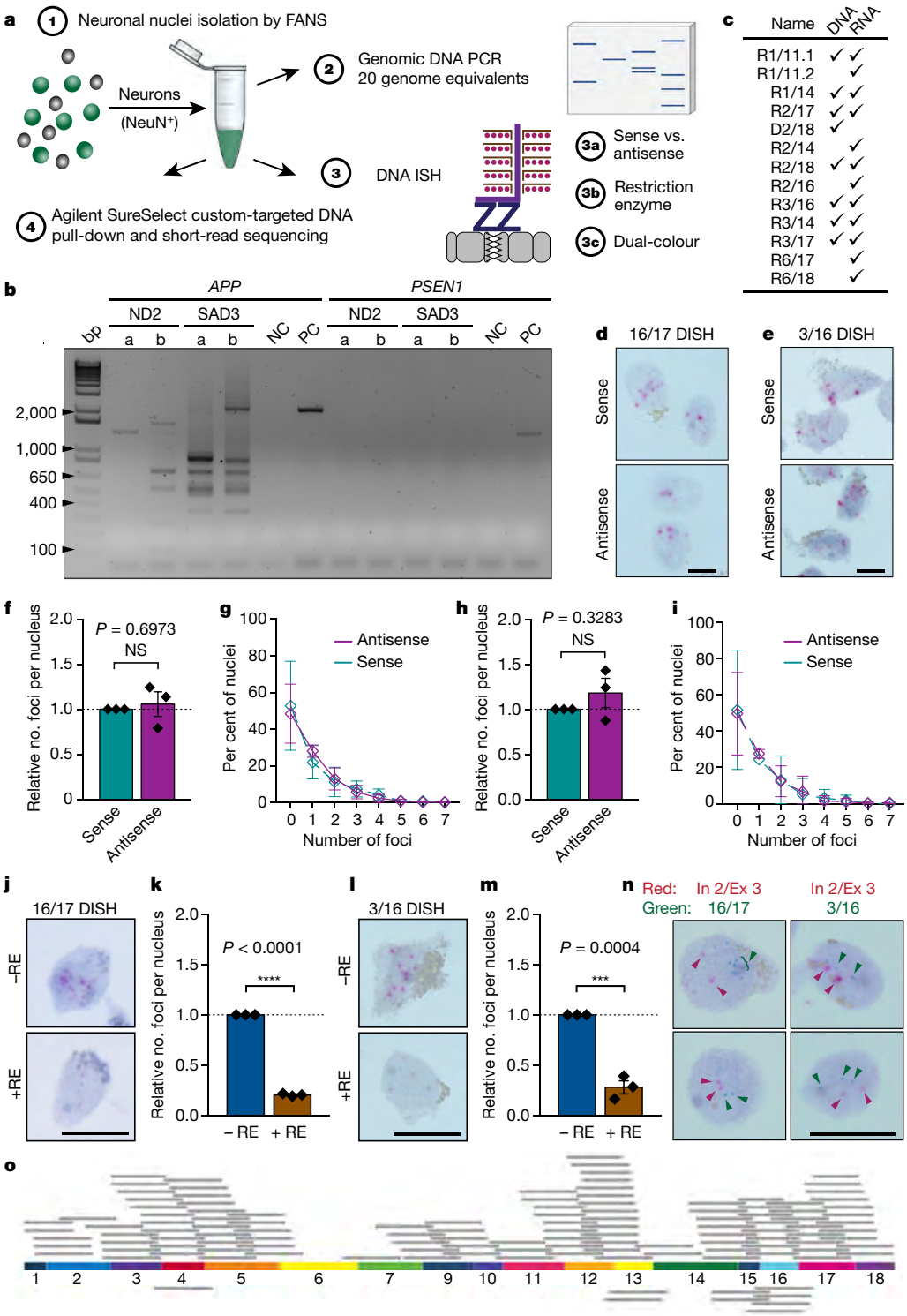


Fig. 2 | *APP* gencDNAs identified by DNA polymerase-dependent and -independent methods. a, FANS-isolated neuronal nuclei from human prefrontal cortices (1) were used for genomic DNA PCR (2), DISH (3), and custom target enrichment followed by deep sequencing (4). **b**, Electrophoresis of genomic DNA PCR products with *APP* and *PSEN1* primer sets, from neurons from normal brains or brains from individuals with SAD with two replicates (a and b). Non-template control (NC) and positive control (PC) with indicated plasmids are shown. **c**, Cloning and Sanger sequencing revealed multiple gencDNA sequences. **d**–**i**, DISH was performed with sense and antisense probes targeting the exonic 16/17 junction (**d**, **f**, **g**; sense $n = 339$ and antisense $n = 335$), and the intra-exonic 3/16 junction (**e**, **h**, **i**; sense $n = 490$, antisense $n = 484$) on neuronal nuclei from individuals with SAD. **f**–**i**, Relative average number of foci (**f**, **h**; points represent average of independent experiments) and frequency distributions (**g**, **i**) showed no significant differences (unpaired, two-tailed Student's *t*-test). **j**–**m**, Restriction enzyme (RE) digestion using MluCI (**j**, **k**) and PstI + MslI (**l**, **m**) to eliminate 16/17 (−RE $n = 349$, +RE $n = 440$) and 3/16 (−RE $n = 367$, +RE $n = 340$) target sequences, respectively. Statistical significance on all bar graphs was determined using unpaired, two-tailed Student's *t*-test. **n**, Dual DISH with intron 2/exon 3 (red) genomic locus and 16/17 or 3/16 probes (green). **o**, Schematic of *APP* cDNA and genomic exon–exon junctions identified by Agilent SureSelect enrichment of the *APP* locus and Illumina sequencing; reads below span two exon–exon junctions. NS, not significant. Error bars show s.e.m. Scale bars, 10 μm.

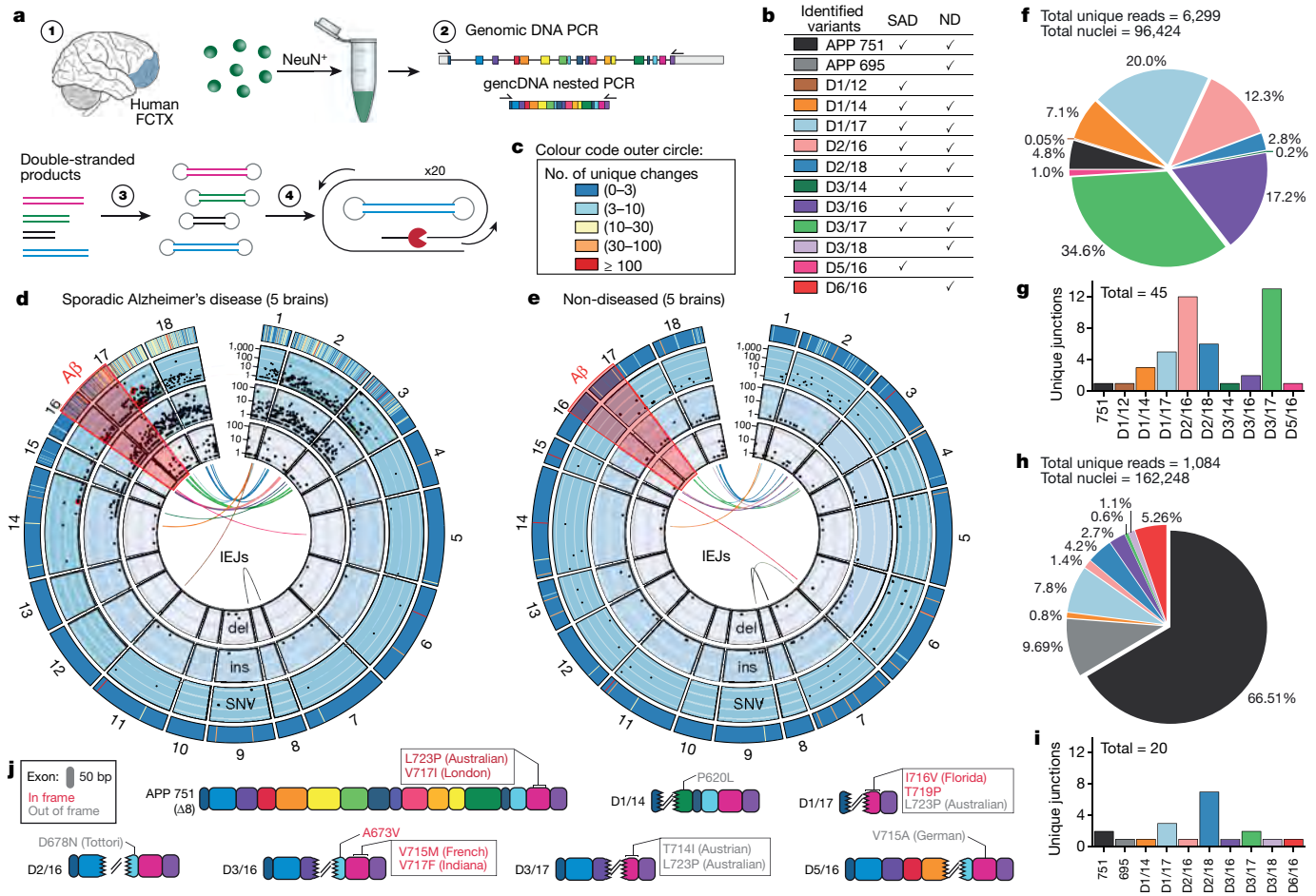


Fig. 3 | Thousands of unique gencDNAs identified by SMRT sequencing of neurons from non-diseased brains and brains of individuals with SAD. a, Neuronal nuclei from prefrontal cortices of individuals with or without SAD were sorted (1) and used for genomic DNA PCR (2). Multiple reactions were pooled for library preparation (3) to enable SMRT-CCS (more than 20 passes) (4). **b**, Exon–exon junctions identified. **c**, Key for outermost circle of **d** and **e**, representing the sum of changes at each genomic location. **d**, **e**, Concentric circle plots of the *APP* locus depicting IEJs (central lines), deletions (del), insertions (ins) and SNVs sequenced from the brains of five individuals with SAD (**d**) and five individuals without SAD (**e**). Black dots indicate the abundance of deletions, insertions and SNVs on a $\log_{10}$ scale at the specified location. The Aβ region is highlighted in red and known FAD-associated mutations are circled in red. **f**, **h**, The percentage of unique reads from each exon–exon junction for brains from individuals with SAD (**f**) and without (**h**). **g**, **i**, Number of unique IEJs from each exon–exon combination for brains from individuals with SAD (**g**) and without (**i**). **j**, Eleven FAD-associated mutations were identified in 6 *APP* gencDNAs and *APP*-751. In-frame (red) and out-of-frame (grey) mutations are based on the known *APP* reading frames.

SureSelect targeted DNA pull-down (Extended Data Fig. 2i), which showed unbiased genomic coverage across the entire *APP* genomic locus (all introns and exons including exon 8; Extended Data Fig. 2j). Analysis of DNA from 40,000 neuronal nuclei from individuals with SAD identified all previously detected gencDNA exon–exon junctions excluding exon 8, which is absent from the brain-specific *APP* mRNA splice variants *APP*-751 and *APP*-695 (Fig. 2o; see also Fig. 1e).

Distinct *APP* gencDNAs in SAD

The diversity of gencDNA sequences was assessed by a distinct technical approach, single molecule real-time (SMRT) circular consensus sequencing (CCS), which enables high-certainty, long-read calls to be produced by multiple passes over the same template. gencDNAs were enriched by multiple PCR reactions on small neuronal populations from five individuals with SAD (149 reactions from 96,424 nuclei) and five healthy brains (244 reactions from 162,248 nuclei; Fig. 3a). Samples were pooled for library preparation and SMRT-CCS. Of note, more non-diseased nuclei than diseased nuclei were required to produce sufficient product for sequencing. We identified 6,299 unique sequences (10% in frame; Extended Data Fig. 3a, b) including 45 different IEJs, in neuronal nuclei from the brains of individuals with SAD, and 1,084 unique sequences (12.1% in frame; Extended Data Fig. 3a, c), including 20 IEJs, in neuronal nuclei from non-diseased brains (Fig. 3b–i).

Critically, both qualitative and quantitative differences in the sequences of gencDNA variants distinguished the brains of individuals with SAD from healthy brains (Fig. 3b–i). Distinctions included gencDNAs with novel IEJs and SNVs (Fig. 3d, e), which were far more prevalent in the brains of individuals with SAD. By contrast, gencDNAs of the canonical neuronal splice variants, *APP*-751 and *APP*-695, predominated in non-diseased brains, and brains from individuals with SAD showed reduced *APP*-751 and no *APP*-695 (Fig. 3f, h). Notably, 11 SNVs that had been previously published as pathogenic FAD mutations (Fig. 3d, j, Supplementary Tables 3, 5), including the Indiana mutation[12], were present in neurons from individuals with SAD. No FAD mutations were detected in non-diseased brains (Fig. 3e, Supplementary Tables 4, 6).

gencDNA formation in cell lines

APP gencDNAs lacking introns, and the presence of brain-specific isoforms (*APP*-751, *APP*-695), support origins of gencDNAs from RNAs that must involve reverse transcription. To model gencDNA production in cell lines, we expressed *APP*-751 cDNA (Fig. 4a) in a Chinese hamster ovary (CHO) cell line with endogenous reverse transcriptase

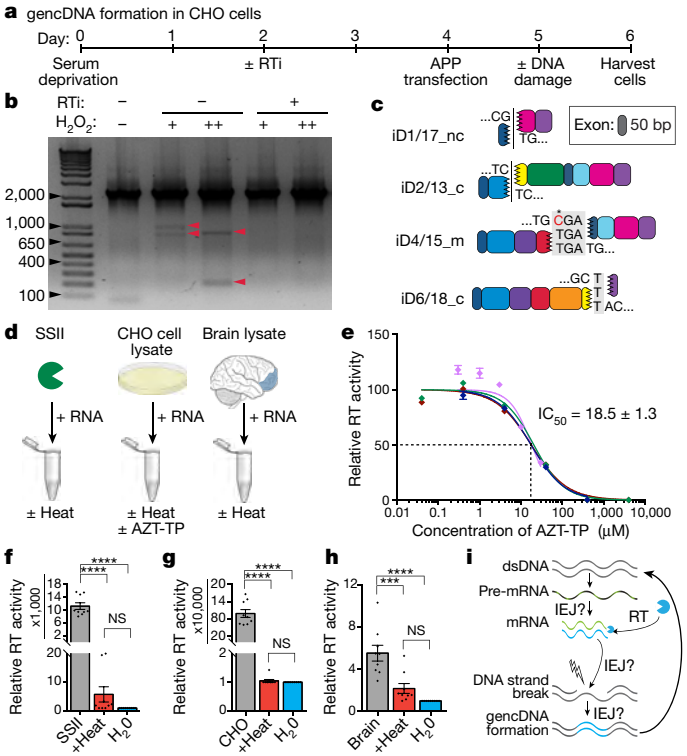


Fig. 4 | Mechanistic studies of gencDNA formation in culture. **a**, Timeline of CHO cell experiments. RTi, AZT (100 μM) and ABC (10 μM); transfection, *APP*-751; DNA damage, H_2O_2 at 5 μM (+) and 50 μM (++). **b**, Gel electrophoresis with red arrowheads indicating cloned and sequenced bands. **c**, New induced gencDNA variants. **d**, Reverse transcriptase (RT) activity was analysed in Super Script II (SSII) positive controls, CHO cell lysate, and human brain lysate. **e**, Four independent experiments showed decreased reverse transcriptase activity in CHO cells in response to the RTi azidothymidine triphosphate (AZT-TP). Colours represent individual experiments. **f**–**h**, Relative reverse transcriptase activity in SSII controls, CHO cells, and brain samples (three independent experiments with three biological replicates). Statistical significance was determined using ordinary one-way ANOVA with Sidak's multiple comparisons test. ***$P = 0.003$, ****$P < 0.0001$. NS, not significant (**f**, **g**, $P > 0.9999$; **h**, $P = 0.3095$). Error bars show s.e.m. **i**, Proposed model of reverse transcriptase activity in the formation of gencDNAs.

activity. Initial results did not show gencDNAs, but induction of DNA strand breaks by H_2O_2 produced novel gencDNAs (Fig. 4b, c). Additionally, endogenous reverse transcriptase activity was also required to produce gencDNAs, based on results using the nucleoside reverse transcriptase inhibitors (RTi) abacavir (ABC) and azidothymidine (AZT) (Fig. 4b). Variant RNAs were also dependent on reverse transcriptase activity (Extended Data Fig. 4). Endogenous reverse transcriptase activity was confirmed in CHO cells and further identified in human prefrontal cortex (Fig. 4d–h), consistent with gencDNA production from RNA intermediates and reverse transcription (Fig. 4i).

Increased gencDNAs in SAD and J20 neurons

We further explored the relationships between gencDNAs and SAD using DISH. We examined two gencDNA junctions, $DISH_{16/17}$ and $DISH_{3/16}$, in neurons from six individuals with verified SAD and six non-diseased brains (Fig. 5a–f, Extended Data Fig. 5a–f, Extended Data Table 2; average age 83.5 and 86.7 years, respectively). The number of red foci in neurons from individuals with SAD was three- to fivefold higher than in non-diseased neurons and ranged from 0 to a maximum of 13 in SAD nuclei. Rare foci were observed in non-neuronal nuclei but were not statistically increased in SAD (Fig. 5a–f, Extended Data Fig. 5a–f). Increased gencDNAs in neurons from individuals with SAD raised the question of whether gencDNAs could give rise to toxic proteins. We therefore tested the cytotoxicity of three *APP* RNA variants (R2/18, R3/14, and R3/16), which were translated in vitro, and found that two of the three variants induced cell death in SH-SY5Y cells (Extended Data Fig. 5g, h).

The J20 mouse model of Alzheimer's disease forms Aβ plaques that accumulate with age. These mice harbour multiple copies of a human *APP* transgene containing the Swedish (K670M/N671L) and Indiana (V717F) mutations, driven by a neuron-specific platelet-derived growth factor-β (PDGF-β) promoter to produce selective, high expression in neurons, with little or no expression in non-neuronal cells[18]. DISH probes for human *APP* did not detect the endogenous mouse locus (Fig. 5g, Extended Data Fig. 6b). $DISH_{3/16}$ identified enriched signals in J20 neuronal nuclei, contrasting with low levels in non-neuronal nuclei from the same mice (Fig. 5g, h, Extended Data Fig. 6a). The more prevalent gencDNA sequence recognized by $DISH_{16/17}$ was also highly enriched in neurons. Notably, $DISH_{16/17}$ demonstrated an age-dependent increase in the area of gencDNA foci over a 2.3-year period, a pattern of change that was not observed in non-neuronal nuclei (Fig. 5i, j, Extended Data Fig. 6c). Use of cells infected with retroviral proviruses containing 0, 1, or 2 copies of the $DISH_{16/17}$ target sequence demonstrated that DISH is semiquantitative and reflects DNA copy numbers (Extended Data Fig. 6d–f). The neuron-selective increase in area of foci occurs during adult life, long after cerebral cortical neurogenesis has ceased[19], further supporting the theory that neuronal gene transcription generates gencDNAs.

Discussion

Human neuronal *APP* gene recombination was identified in brains from healthy controls and individuals with SAD. It was characterized by the mosaic presence of thousands of distinct gencDNA variants that enter neuronal genomic DNA through a process involving *APP* transcription that is influenced by neural activity, DNA strand breaks and reverse transcription (Supplementary Discussion). *APP* gencDNAs bear some resemblance to, but are fundamentally distinct from, processed pseudogenes[20] (non-coding, germline remnants of evolutionarily retrotransposed mRNAs[21,22] that can be active in cancers[23]) and LINE1 repeat elements (which encode an active reverse transcriptase (ORF2)[20] to allow potential retrotransposition in mitotic cells, including within the developing brain[5,9,24–26]). By comparison, *APP* gencDNAs manifest as thousands of distinct genomic variants derived from a cellular gene, contain IEJs and myriad SNVs, can undergo multiple 'retro-insertions' into post-mitotic neuronal genomes, and appear capable of being actively transcribed and translated to produce variant bioactive products that are relevant to both normal and diseased states.

Constitutive mutations or *APP* CNVs are considered causal in FAD and Down syndrome, raising the possibility that previously reported somatic *APP* exonic CNVs contribute mechanistically to SAD[10], which can be explained by the somatic gene recombination identified here. Proof-of-concept data from individuals with SAD identified a marked shift in the forms and abundance of gencDNAs when compared with healthy controls (Figs. 3, 5), including the three- to fivefold increase in gencDNAs in all brains from individuals with SAD examined (Fig. 5). Notably, we identified 11 somatic SNVs that were previously identified as being pathogenic in FAD[12], which were absent from non-diseased controls. Other SNVs, as well as myriad genomic alterations, may also contribute to SAD through both classical and non-classical mechanisms.

Classical mechanisms that support the amyloid hypothesis involve production of toxic Aβ peptides and plaque formation. gencDNAs, including those with FAD-associated mutations, are likely to represent a source of secretase-cleaved substrate for the production of Aβ, as well as a potential source of toxic products that do not require secretase cleavage (Extended Data Fig. 5). Non-ATG translation could potentially occur for out-of-frame variants[27]. Non-classical mechanisms might involve RNA pathologies[28] or maximum limits to gencDNA integration per neuron, beyond which neurodegeneration occurs via genome instability, akin to deleterious mobile elements[29]. The potential diversity of

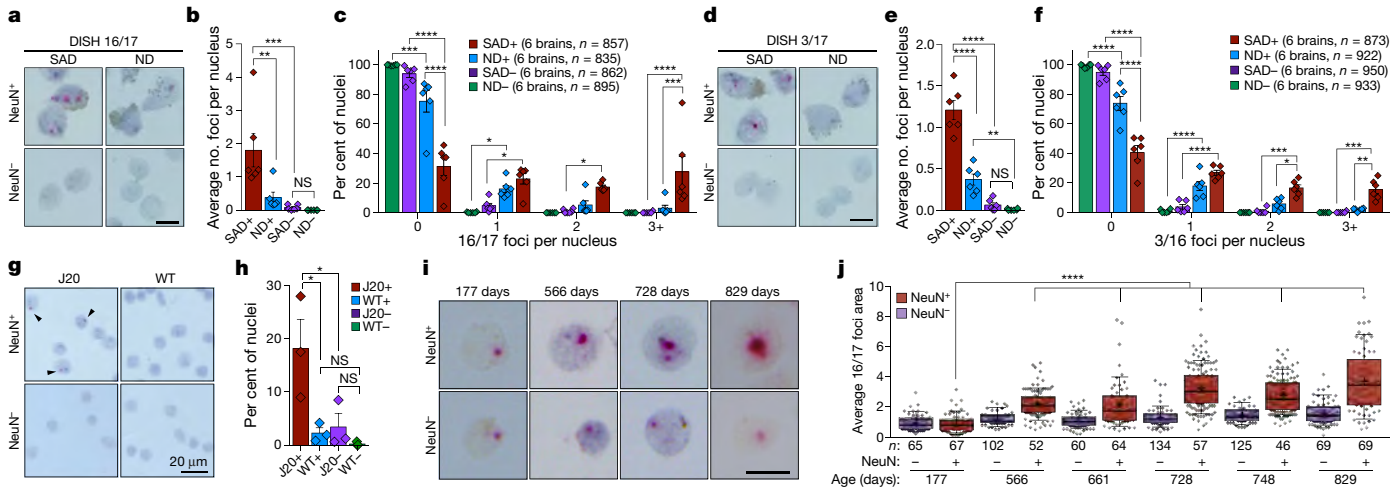


Fig. 5 | Proof-of-concept correlation between gencDNAs and SAD. **a**–**f**, Nuclei sorted from the brains of six individuals with SAD and six individuals without SAD were analysed by $DISH_{16/17}$ (**a**–**c**) and $DISH_{3/16}$ (**d**–**f**). **a**, **d**, Representative DISH images. **b**, **e**, Average number of foci per nucleus was increased in neurons from individuals with SAD (one-way ANOVA with Holm–Sidak's multiple comparison test). **c**, **f**, Frequency distributions displaying the percentage of nuclei with 0, 1, 2 and 3 or more (3+) foci (two-way ANOVA with Tukey's multiple comparison test). **g**, Representative $DISH_{3/16}$ of J20 and wild-type (WT) nuclei. **h**, The percentage of nuclei with one or more foci was increased in J20 neurons (ordinary one-way ANOVA with Sidak's multiple comparisons test, J20+ versus WT+ $P = 0.0253$, J20+ versus J20− $P = 0.0371$, WT+ versus WT− $P = 0.9267$, J20− versus WT− $P = 0.9842$). **i**, **j**, Area of $DISH_{16/17}$ foci increased with age in J20 mice. **i**, Representative images of from mice aged 177, 566, 661, 728, 748 and 829 days (one animal each, number of nuclei interrogated is listed below box). **j**, Area of foci shows statistically significant increases with age. +, mean; line, median; box, 75th–25th percentiles; whiskers, 90th–10th percentiles (non-parametric Kruskal–Wallis with Dunn's multiple comparisons test). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. NS, not significant. Detailed *P* values for **b**, **c**, **e**, **f**, **j** are listed in Extended Data Figs. 5, 6. Error bars show s.e.m. Scale bars, 10 μm unless otherwise noted.

protein variants produced by gencDNAs, and other non-protein mechanisms, may help to explain the failure of therapeutic trials targeting Aβ and related enzymologies, especially those targeting single molecular entities[30].

The largest risk factor for SAD is age, and the age-related increase in gencDNA variants in neurons offers a possible explanation for the decades of life required for SAD to manifest. Neuronal *APP* transcription promotes gencDNA generation both in cell culture and in J20 neurons in vivo (Figs. 4, 5). This is consistent with the increase in *APP* transcription that was previously linked to SAD incidence[31], and the gene encoding the SAD risk factor ApoE4[32]. Notably, the dependence of gencDNA on reverse transcriptase activity might be relevant to the statistical rarity of proven cases of Alzheimer's disease in individuals with HIV infection who are more than 65 years old[33,34], and who have received prolonged, combined anti-retroviral therapy (cART), which includes reverse transcriptase inhibitors. If confirmed, this observation would suggest the immediate use of FDA-approved cARTs or modified combinations containing reverse transcriptase inhibitors to treat SAD, Down syndrome and perhaps FAD. Additionally, processes that produce DNA breaks, such as head injury, that have been linked to Alzheimer's disease[35–37] are consistent with gencDNA production requiring DNA breaks. Thus, gencDNAs and their production have properties relevant to a range of Alzheimer's disease mechanisms and the development of new therapeutic strategies.

The presence of *APP* gencDNAs in non-diseased neurons is likely to reflect the normal roles of *APP*[38], including synaptic function; here, *APP* gencDNAs might provide an increased repertoire of protein species, contributing to synaptic diversity. Additional genes may be transcriptionally modified and genomically retro-inserted in response to selective activities in neuronal populations. Such a mechanism might enable preferential gene re-expression that bypasses splicing or further RNA modification. More broadly, gencDNAs could provide neurons with an activity-dependent mechanism for recording and retaining information over long periods of time, perhaps placing multiple forms of a gene under transcriptional control distinct from a wild-type locus, which could be produced through diverse genomic integration sites that remain to be determined. Such a process could have relevance to known neuronal functions that depend on transcriptional activity, including Hebbian plasticity[39], synaptic wiring[40], learning and memory[41], and cognition[42]. Thus, gencDNA production may represent both a 'recording' and a 'playback' mechanism for expressing a symphony of variants beyond wild-type gene forms. It would be surprising if *APP* were the only gene to undergo this form of recombination, which might influence distinct, normal brain functions as well as contributing to brain disorders such as Alzheimer's disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0718-6.

Received: 3 October 2017; Accepted: 9 October 2018;
Published online: 21 November 2018

1. Dreyer, W. J., Gray, W. R. & Hood, L. The genetic, molecular and cellular basis of antibody formation: some facts and a unifying hypothesis. *Cold Spring Harb. Symp. Quant. Biol.* **32**, 353–367 (1967).
2. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl Acad. Sci. USA* **73**, 3628–3632 (1976).
3. Chun, J. J., Schatz, D. G., Oettinger, M. A., Jaenisch, R. & Baltimore, D. The recombination activating gene-1 (RAG-1) transcript is present in the murine central nervous system. *Cell* **64**, 189–200 (1991).
4. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).
5. Rohrback, S., Siddoway, B., Liu, C. S. & Chun, J. Genomic mosaicism in the developing and adult brain. *Dev. Neurobiol*. https://doi.org/10.1002/dneu.22626 (2018).
6. Rehen, S. K. et al. Chromosomal variation in neurons of the developing and adult mammalian nervous system. *Proc. Natl Acad. Sci. USA* **98**, 13361–13366 (2001).
7. Rehen, S. K. et al. Constitutional aneuploidy in the normal human brain. *J. Neurosci.* **25**, 2176–2180 (2005).
8. Westra, J. W. et al. Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. *J. Comp. Neurol.* **518**, 3981–4000 (2010).
9. McConnell, M. J. et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* **356**, eaal1641 (2017).
10. 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).
11. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8**, 595–608 (2016).

12. Murrell, J., Farlow, M., Ghetti, B. & Benson, M. D. A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science* **254**, 97–99 (1991).
13. Hooli, B. V. et al. Rare autosomal copy number variations in early-onset familial Alzheimer's disease. *Mol. Psychiatry* **19**, 676–681 (2014).
14. Wiseman, F. K. et al. A genetic cause of Alzheimer disease: mechanistic insights from Down syndrome. *Nat. Rev. Neurosci.* **16**, 564–574 (2015).
15. Rohrback, S. et al. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. https://doi.org/10.1073/pnas.1812702115 *Proc. Natl Acad. Sci. USA* (2018).
16. Lake, B. B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
17. Dawkins, E. & Small, D. H. Insights into the physiological function of the β-amyloid precursor protein: beyond Alzheimer's disease. *J. Neurochem.* **129**, 756–769 (2014).
18. Mucke, L. et al. High-level neuronal expression of $A\beta_{1-42}$ in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. *J. Neurosci.* **20**, 4050–4058 (2000).
19. Ming, G. L. & Song, H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* **70**, 687–702 (2011).
20. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000).
21. Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N. & Gerstein, M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res.* **33**, 2374–2383 (2005).
22. Vanin, E. F. Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19**, 253–272 (1985).
23. Kalyana-Sundaram, S. et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149**, 1622–1634 (2012).
24. 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).
25. Muotri, A. R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
26. Upton, K. R. et al. Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228–239 (2015).
27. Cleary, J. D. & Ranum, L. P. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. *Curr. Opin. Genet. Dev.* **26**, 6–15 (2014).
28. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
29. McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801 (1984).
30. Egan, M. F. et al. Randomized trial of verubecestat for mild-to-moderate Alzheimer's disease. *N. Engl. J. Med.* **378**, 1691–1703 (2018).
31. Brouwers, N. et al. Genetic risk and transcriptional variability of amyloid precursor protein in Alzheimer's disease. *Brain* **129**, 2984–2991 (2006).
32. Huang, Y. A., Zhou, B., Wernig, M. & Sudhof, T. C. ApoE2, ApoE3, and ApoE4 differentially stimulate APP transcription and Aβ secretion. *Cell* **168**, 427–441. e21 (2017).
33. Turner, R. S. et al. An individual with human immunodeficiency virus, dementia, and central nervous system amyloid deposition. *Alzheimers Dement. (Amst.)* **4**, 1–5 (2016).
34. Centers for Disease Control and Prevention. *HIV Surveillance Report, 2016* Vol. 28 http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html (2017).
35. Suberbielle, E. et al. Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-β. *Nat. Neurosci.* **16**, 613–621 (2013).
36. Mortimer, J. A., French, L. R., Hutton, J. T. & Schuman, L. M. Head injury as a risk factor for Alzheimer's disease. *Neurology* **35**, 264–267 (1985).
37. Pravdenkova, S. V., Basnakian, A. G., James, S. J. & Andersen, B. J. DNA fragmentation and nuclear endonuclease activity in rat brain after severe closed head injury. *Brain Res.* **729**, 151–155 (1996).
38. Nhan, H. S., Chiang, K. & Koo, E. H. The multifaceted nature of amyloid precursor protein and its proteolytic fragments: friends and foes. *Acta Neuropathol.* **129**, 1–19 (2015).
39. Guzman-Karlsson, M. C., Meadows, J. P., Gavin, C. F., Hablitz, J. J. & Sweatt, J. D. Transcriptional and epigenetic regulation of Hebbian and non-Hebbian plasticity. *Neuropharmacology* **80**, 3–17 (2014).
40. Hattori, D., Millard, S. S., Wojtowicz, W. M. & Zipursky, S. L. Dscam-mediated cell recognition regulates neural circuit formation. *Annu. Rev. Cell Dev. Biol.* **24**, 597–620 (2008).
41. Madabhushi, R. et al. Activity-induced DNA breaks govern the expression of neuronal early-response genes. *Cell* **161**, 1592–1605 (2015).
42. West, A. E. & Greenberg, M. E. Neuronal activity-regulated gene transcription in synapse development and cognitive function. *Cold Spring Harb. Perspect. Biol.* **3**, a005744 (2011).

Acknowledgements We thank D. Schatz, C. Murre and J.-P. Changeux for discussions; the UCSD ADRC and the UCI MIND for human brain specimens, along with the donors and families who shared these precious materials; flow cytometry core colleagues B. Seegers, M. Haynes (TSRI) and Y. Altman (SBP); and M. Wang (ACD), D. J. Weiss (Agilent) and H. Lee (PacBio) for technical assistance. Support was provided by The Shaffer Family Foundation, The Bruce Ford & Anne Smith Bundy Foundation, a UCSD pilot grant (NIH P50AG00513) and SBP institutional funds (J.C.); a PRAP fellowship from the Ministry of Science and Technology, Taiwan (105-2917-I-564-085, M.-H.L.); and NIH training grant 5T32AG000216-24 (G.E.K.).

Reviewer information *Nature* thanks L. Feuk, F. Gage and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.C. conceived the project. M.-H.L. and J.C. designed and analysed experiments. M.-H.L. (nuclei sorting, RT–PCR, Southern blot, cloning and Sanger sequencing, genomic DNA PCR, RNA and DNA in situ hybridization, targeted DNA pull-down and Illumina sequencing, SMRT sequencing, gencDNA induction in culture and cytotoxicity assay), B.S. (SMRT sequencing), R.R. (genotyping, breeding and maintenance of mouse lines and DNA concatemer preparation and virus infection), W.J.R. (nucleus sorting, in vitro reverse transcriptase activity assay and western blot), C.P. (genomic DNA PCR) and G.K. (RNA and DNA in situ hybridization) completed experiments. M.-H.L. and G.E.K. performed statistical analyses and made figures. B.S., I.S., C.S.L. and T.L. performed informatics analyses. M.-H.L., B.S., G.E.K. and J.C. prepared the manuscript. Key experiments were repeated by G.E.K., W.J.R., C.P. and other researchers in our laboratory.

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.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0718-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0718-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.C.

METHODS

Human brain tissue and J20 mice. Fresh frozen human brain tissue was provided by the University of California San Diego (UCSD) Alzheimer's Disease Research Center (ADRC) and the University of California Irvine (UCI) Institute for Mind Impairments and Neurological Disorders (MIND).

J20 transgenic mice (B6/Cg-Tg(PDGFB-APPSwInd)20Lms/2JMmjax) were purchased from The Jackson Laboratory and housed in IACUC approved animal facilities in accordance with applicable laws and regulations at Sanford Burnham Prebys Medical Discovery Institute. Sex (F/M) and age (days) of mice used for experiments are listed: F177, M566, M661, M661, M728, F748, F829, and M861. Sample sizes were estimated based upon preliminary data without additional statistics. Samples were allocated randomly and in situ hybridization quantification was blinded for statistical assessments.

Nucleus extraction and FANS. Human and mouse brain nuclei were isolated as described previously[10]. For in situ hybridization analyses, isolated nuclei were fixed in 1:10 diluted buffered formalin (Fisher Healthcare) for 5 min. Prior to sorting, fixed or unfixed nuclei were then labelled with anti-NeuN rabbit monoclonal antibody (1:800) (Millipore, Germany) and Alexa Fluor 488 donkey anti-rabbit IgG (1:500) (Life Technologies, Carlsbad, CA), and counterstained with propidium iodide (PI; 50 μg/ml) (Sigma, St. Louis, MO). For DNA analyses, RNase A (100 μg/ml) was included with all subsequent steps after initial nuclei isolation, including primary and secondary antibody incubations. Diploid NeuN-positive and negative nuclei were gated by PI and immunofluorescence, and sorted into appropriate populations for RT–PCR, genomic DNA PCR, or in situ hybridization. FANS was performed on a FACSAria Fusion (BD Biosciences, Franklin Lakes, NJ) or with a FACS-Aria II (BD Biosciences, Franklin Lakes, NJ).

RNA extraction and RT–PCR. All RNA extractions from 50-nuclei populations (NeuN-negative and positive) and bulk tissues were performed using Quick-RNA MicroPrep (Zymo Research, Irvine, CA) and RNAeasy Mini kits (Qiagen, Valencia, CA) according to the manufacturer's protocols. OneStep Ahead RT–PCR (Qiagen, Valencia, CA) was used for RT–PCR with *APP* 1–18 primer sets (Supplementary Table 1) according to the manufacturer's protocol. Oligo-$(dT)_{20}$ primer was used to prime the cDNA library as indicated. Low annealing stringency PCR was carried out with the following thermal cycling steps for 40 cycles: 95 °C 15 s, 55 °C for *APP* and 52 °C for *PSEN1* 15 s, and 68 °C 2.5 min for *APP* and 2 min for *PSEN1*.

Southern blotting. RT–PCR products were run on agarose gel, denatured, and transferred to a positively charged nylon membrane. UV-crosslinked membranes were incubated with denatured and purified ^{32}P-labelled *APP* cDNA probes at 42 °C overnight. Blots were washed four times with increasing washing stringency and temperature according to established protocols. Images were developed on a Typhoon (GE Healthcare Life Sciences) or Fujifilm FLA-5100 phosphorimager.

DNA extraction and genomic DNA PCR. DNA extraction from isolated neuronal nuclei populations was performed via isopropanol precipitation. In brief, nuclei were incubated with proteinase K in 550 μl PK buffer (50 mM Tris pH 8.0, 0.1 M EDTA, 0.1 M NaCl, 1% SDS) overnight at 55 °C. Samples were then treated with RNase cocktail enzyme mix (ThermoFisher, Waltham, MA) for 2 h, followed by addition of 250 μl saturated NaCl. After centrifugation, supernatant was used for DNA precipitation by isopropanol and washed three times with 70% ethanol. DNAeasy and QIAamp DNA Mini kits (Qiagen, Valencia, CA) were also used according to the manufacturer's instructions. Purified DNA was stored at −20 °C for future use. High annealing stringency PCR was performed using either the FastStart PCR master mix (Sigma, St. Louis, MO) with PCR cycle settings: 95 °C 30 s, 65 °C for 30 s, and 72 °C 2.5 min for 40 cycles, or the Platinum SuperFi DNA polymerase (ThermoFisher, Waltham, MA) with cycle settings: 98 °C 10 s, 65 °C for *APP* and 52 °C for *PSEN1* for 10 s, and 72 °C 1.5 min for *APP* and 1 min for *PSEN1* for 40 cycles. Primer sequences are listed in Supplementary Table 1.

DNA in situ hybridization and RNA in situ hybridization. For DISH pretreatment, sorted nuclei were dried on Plus Gold slides (Fisher Scientific, Pittsburgh, PA). Nuclei were then treated with RNase cocktail enzyme mix (RNase A + RNase T1, 1:50) (ThermoFisher, Waltham, MA) at 40 °C for 60 min, followed by fixation in 1:10 dilution buffered formalin at room temperature for 5 min. After being washed with distilled water twice, slides were treated with hydrogen peroxide at room temperature for 10 min, target retrieval reagent at 95 °C for 15 min, followed by protease treatment at 40 °C for 10 min. Restriction enzyme was applied after protease treatment for 2 h as necessary in negative control experiments. DNA was then denatured (2×SSC, 70% formamide and 0.1% sodium dodecyl sulfate) at 80 °C for 20 min. After cooling the slides to room temperature, BaseScope probes were applied and incubated with nuclei at 40 °C overnight. Samples were then prepared for signal development. For RISH pretreatment, 10 μm fresh frozen human tissue sections was fixed in 1:10 dilution buffered formalin on ice for 10 min. After washing with PBS twice, tissue sections were placed in serial diluted ethanol 50%, 70% and 100%, 5 min for each step. Slides were then treated with hydrogen peroxide at room temperature for 10 min, followed by protease treatment at room temperature for 20 min. BaseScope probes were then incubated with tissue sections at 40 °C for 2 h. Hydrogen peroxide, 10× target retrieval buffer, proteases, custom BaseScope probes (Supplementary Table 1), and BaseScope reagent kit-RED used for signal development were all purchased from Advanced Cell Diagnosis (ACD, Newark, CA). Duplex BaseScope reagent kit was also purchased from ACD. Nuclei and tissue sections were counterstained with haematoxylin. Zeiss AX10 Imager, M2 microscope and ZEN2 software were used for image acquisition. Images were thresholded, and foci number or size were quantified using ImageJ for statistical analysis.

Agilent SureSelect hybridization enrichment and sequencing. The method is graphically represented in Extended Data Fig. 2i. Nuclei were isolated from human frontal cortex, labelled for NeuN, and NeuN-positive nuclei were isolated via FANS. Genomic DNA was extracted and fragmented into ~1.2 kb using sonication (Covaris, Woburn, MA). End repair reactions were performed and Illumina sequencing adaptors were ligated to genomic DNA. Library-prepped DNA was hybridized with custom Agilent SureSelect probes designed against the entire *APP* locus, including introns. Purified *APP*-containing genomic DNA sequences were then sequenced on an Illumina NextSeq (Illumina, San Diego, CA). Sequences were aligned to the human reference genome (GRCh38) using STAR (version 2.5.3a) with the settings: --outSAMattributes All --outFilterScoreMinOverLread 0.8 --outSJfilterCountTotalMin 1 1 1 1. Duplicate reads were marked and removed using Picard (version 2.1.1). Reads were then informatically analysed using IGV, the UCSC Genome Browser, and a custom imaging pipeline built in R.

SMRT sequencing. Neuronal genomic DNA was isolated as described above and used for *APP* PCR and nested *APP* PCR. Platinum SuperFi DNA polymerase with 100× higher fidelity compared to native *Taq* (Invitrogen, Platinum SuperFi DNA Polymerase) was used under high annealing stringency (98 °C 10 s, 65 °C 10 s, and 72 °C 1.5 min, for 30 cycles). An aliquot of the first PCR product was used as a DNA template for nested PCR reactions. Multiple PCR reactions were pooled (149 reactions for Alzheimer's disease and 244 reactions for non-diseased) and purified by DNA Clean and Concentrator-5 (Zymo Research, Irvine, CA) for SMRT sequencing library preparation. PCR amplicons were repaired using SMRTbell template prep kit version 2.0 (PacBio) and purified using AMPure PB beads (PacBio, Menlo Park, CA). Adapters were ligated to DNA to create SMRTbell libraries. Sequencing polymerase was annealed and the SMRTbell library was loaded using Magbead binding. Raw bam sequencing files were converted to fastq format using the CCS algorithm in the SMRTLink software tool kit from PacBio. In CCS, reads were included in the fastq file only if 1) there were more than 20 passes of the sequencing polymerase over the DNA insert in the zero mode waveguide well and 2) the predicted accuracy in SMRTLink was calculated to be greater than 0.9999. These cutoffs generated ultra-high accuracy reads, the median Phred score of reads used was 93[43–46], representing 99.999999% accuracy, with further quality filtering steps applied in our informatic analysis. This SMRT sequencing is comparable in fidelity to Sanger sequencing[43,45].

Genomic data analyses with customized bioinformatic pipelines. Novel algorithms were developed to detect and analyse exon rearrangement in genes of interest. The algorithms were specifically designed to analyse long-read sequences generated by the Pacific Biosciences Sequel platform. A series of quality control (QC) procedures were performed before sequence processing to ensure high quality of the reads being analysed.

Quality control: consensus sequence and read quality. PacBio circular consensus sequence (CCS) reads with fewer than 20 passes were filtered out to ensure overall sequence quality. Quality score distributions were examined: for *APP* gene PCR enriched sequences, average median read-wide Phred score was 93. In gencDNA analyses, we included only reads that met a mean Phred cutoff of >85.

Quality control: sequencing artefacts. Owing to the intrinsic limits of PacBio SMRT sequencing technology, errors in homopolymers (that is, sequence ATTTG could be read as ATTTTG or ATTG in addition to ATTTG) are specially handled with a method that combines quality score information and reference sequence at the beginning of the homopolymer. The FASTQ files encoded uncertainty in the homopolymer run length in the first Phred score of each run. If this Phred score was lower than our threshold of 30, then this position was marked as a likely sequencing artefact and not a real variant.

PCR primer filter. The reads were checked to ensure the correct start and end sites with forward and reverse PCR primer sequences. BLAST (command line tool 'blastn' 2.6.0+) was used to align primer sequences in either orientation to each read with word size 13, gap open penalty 0, and gap extension penalty 2. Any read in which both primers were not detected was filtered out. Furthermore, reads on the negative strand were reverse complemented in this step. BLAST seed length was optimized to avoid ambiguity and ensure sensitivity.

Alignment to APP reference sequences. The Ensembl reference sequence for *APP* was downloaded from the GRCh38 reference human genome assembly using the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) with RefSeq accession number NM_000484.3. Because the PCR primers begin at the start codon and end with the stop codon, sequences of exons 1 and 18 were trimmed to these

positions so that only the coding sequence of each of the 18 exons was kept and stored as a FASTA file. Then, we used BLAST to look for local alignment between 18 exons and each quality-filtered CCS read; blastn parameters used: –outfmt 6, -wordsize 25, -gapopen 0, -gapextend 2. We used the resulting alignment coordinates to mark regions of each read covered by exons. This allowed us to analyse exon arrangements, lengths and patterns of exon–exon junctions.

SNV and INDEL analysis. First, we used reference sequences of *APP* exons to replace low quality individual nucleotides (potential homopolymer runs and other errors) within each read with their reference *APP* exon counterpart. Then, we analysed BLAST local alignments between each exon (or part of an exon) and the read sequence, nucleotide by nucleotide, to look for alignment mismatches. If the mismatch position was a different nucleotide, we assigned it as a single nucleotide variant (SNV); if the mismatch position was a hyphen in the exon sequence, we assigned it as an insertion and if the mismatch position was a hyphen in the read sequence we assigned it as a deletion.

gencDNA production in culture. The method is graphically represented in Fig. 4a. CHO cells were serum-deprived for 2 days, followed by addition of reverse transcriptase inhibitors AZT (100 μM) and ABC (10 μM) (Tocris, Minneapolis, MN) until the end of the experiment. The medium was changed daily with fresh reverse transcriptase inhibitors. Cells were transfected with *APP*-751 driven by the *CAG* promoter by GenJet (SignaGen Laboratories, Gaithersburg, MD) on day 3, then on day 4, cells were treated with 0 μM, 5 μM, or 50 μM hydrogen peroxide (Fisher Scientific) for 2 h. After 1 day, cells were collected and genomic DNA was extracted for PCR analysis.

In vitro reverse transcriptase activity assay. Lysates were prepared in reverse transcriptase disruption buffer[47], and contained cOmplete, EDTA-free protease inhibitor cocktail (Sigma-Aldrich, St. Louis, MO) and PhosSTOP phosphatase inhibitors (Sigma-Aldrich, St. Louis, MO). The assay was performed essentially as described[48], except the assay was separated into two parts. One microgram of extract was used in the reverse transcription step of the assay. In addition, Primer A was used in the reverse transcription reaction cocktail instead of Primer B (Supplementary Information1). Reverse transcription was carried out at 37 °C for 45 min, followed by 15 min at 70 °C.

The reverse transcriptase product of this first step was assayed in triplicate by quantitative PCR. Levels of reverse transcription activity were determined by the Delta Cq method, compared to the activity in negative controls (water and no nucleotides), which were given Cq scores of 40. 100,000 picounits of SuperScript II Reverse Transcriptase (ThermoFisher Scientific) were used as a positive control for the assay.

Lysates for heat inactivation experiments were incubated for 15 min at 70 °C before the reverse transcription step. For inhibitor experiments, lysates were incubated with inhibitor in the presence of all the components of the reaction except for dNTPs. After 10 min at room temperature, dNTPs were added and the reaction was incubated at 37 °C as above. AZT-TP was purchased from TriLink Biotechnologies (San Diego, CA).

Construction and retroviral transduction of synthetic human *APP* sequence targets. Phosphorylated oligonucleotides (Integrated DNA Technologies) composed of human *APP* target sequences with BamHI and BglII restriction sites on the 5′ ends were annealed and ligated into the BamHI site of the retroviral expression vector S-003-AB LZRSpBMN-linker-IRES-EGFP. All primer sequences for construction are listed in Supplementary Table 1. Single and concatamerized oligonucleotide inserts were identified by PCR using primers flanking the BamHI insertion site, and identified clones were sequenced to confirm insert copy number (Genewiz, La Jolla, CA). Helper-free ecotropic virus was produced by transfecting DNA constructs (Lipofectamine 2000, Thermo Fisher Scientific) with single or multiple copies of the oligonucleotide inserts into the retrovirus packaging line Phoenix-ECO. Forty-eight hours after transfection, retroviral supernatants were harvested, and 2 ml of selected virus was used for transduction of NIH-3T3 cells in 6-well plates. Retroviral transduction was carried out by removing the cell growth medium, replacing it with 2 ml retroviral supernatant containing 4 μg/ml polybrene, and spinning at 25 °C for 1 h at 2,800 r.p.m. Forty-eight hours after transduction, the percentage of GFP^+ cells, as identified by flow cytometry, was used to evaluate transduction efficiency.

Cell culture. NIH-3T3 (DMEM, 5% FBS), CHO-K1 (RPMI, 10% FBS), SH-SY5Y (DMEM/F12, 10% FBS), IMR-90 (DMEM, 10% FBS) and HEK-293 (DMEM, 10% FBS) cells were purchased from ATCC and maintained at 37 °C under 5% CO_2. Although HEK-293 is in the database of commonly misidentified cell lines, this cell line was used to show protein expression of our targets in mammalian cells and as gencDNA negative controls. For the indicated purpose, there is no concern regarding this cell line in the manuscript. Cells from ATCC and all reagents used were verified to be mycoplasma free.

Western blot. Cells were harvested in RIPA buffer (100 mM Tris–HCl, pH 7.6, 250 mM NaCl, 0.1% sodium dodecyl sulfate, 0.2% deoxycholic acid, 0.5 mM dithiothreitol, 1 mM EDTA, 0.5% NP-40 and 1% Triton X-100), and proteins in each lysate were analysed using rat monoclonal anti-HA antibody (Clone 3F10, Roche) and horseradish peroxidase-conjugated goat anti-rat secondary antibody (Cell Signaling). Enhanced chemiluminescent substrate (Millipore) for the reaction was added, and the signal was detected by BioRad bioimaging system.

Variant toxicity assay. SH-SY5Y cells were transfected with *APP* RNA variants by lipofectamine LTX (Life Technologies) overnight and further cultured under serum-deprived conditions for 7 days. Cell viability was determined by WST1 reagent (Roche Applied Science) according to the manufacturer's protocol. In brief, cells in 96-well plates were incubated with 100 μl WST1 reagent and culture medium in a ratio of 1:10 (v/v) per well at 37 °C for 2 h. The absorbance at 440 nm of samples normalized by a background control was measured by microplate ELISA reader.

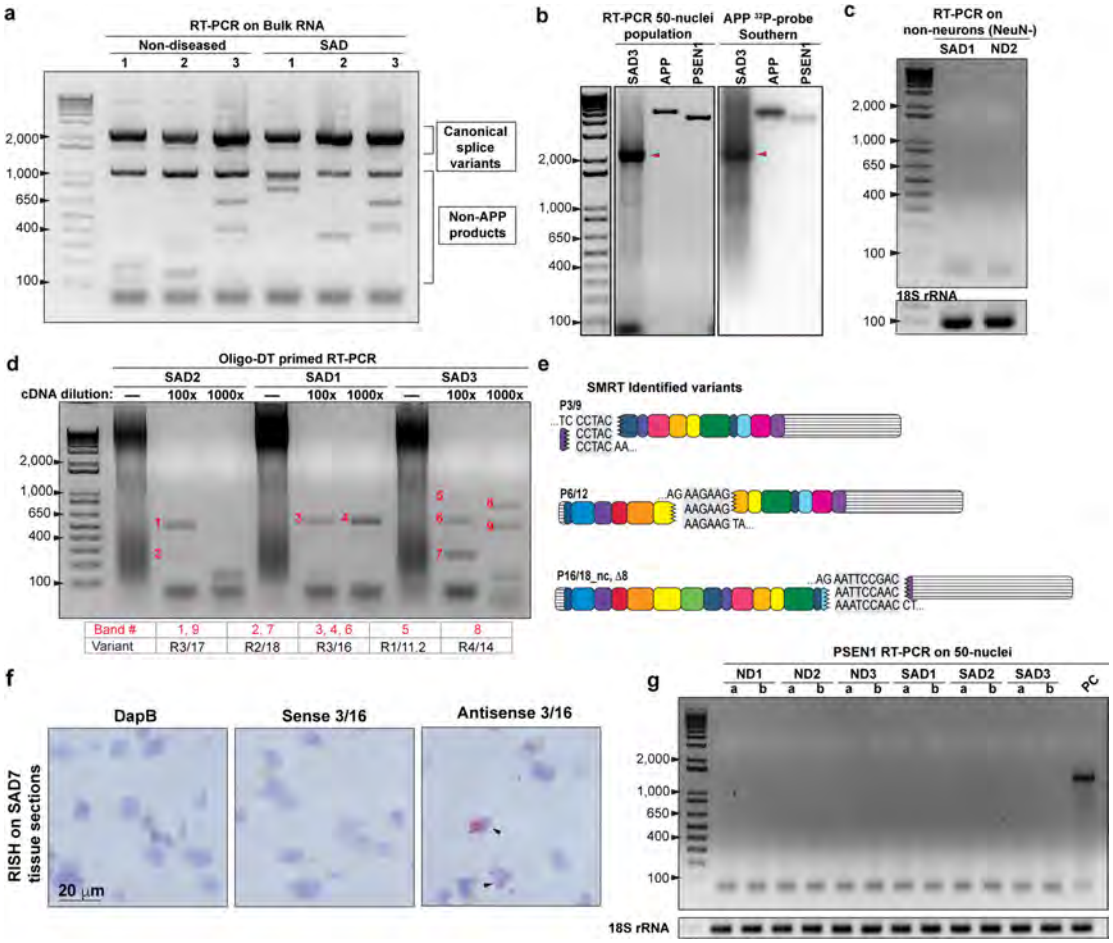
Statistics and reproducibility. All statistical analyses were completed using Prism Version 7 (GraphPad). Specific tests, number of data points (*n*), and *P* values are reported in Figures, figure legends or Extended Data Figures. All experiments in Figures and Extended Data Figures were repeated at least three times (independent experiments) unless specified otherwise in the figure legends.

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

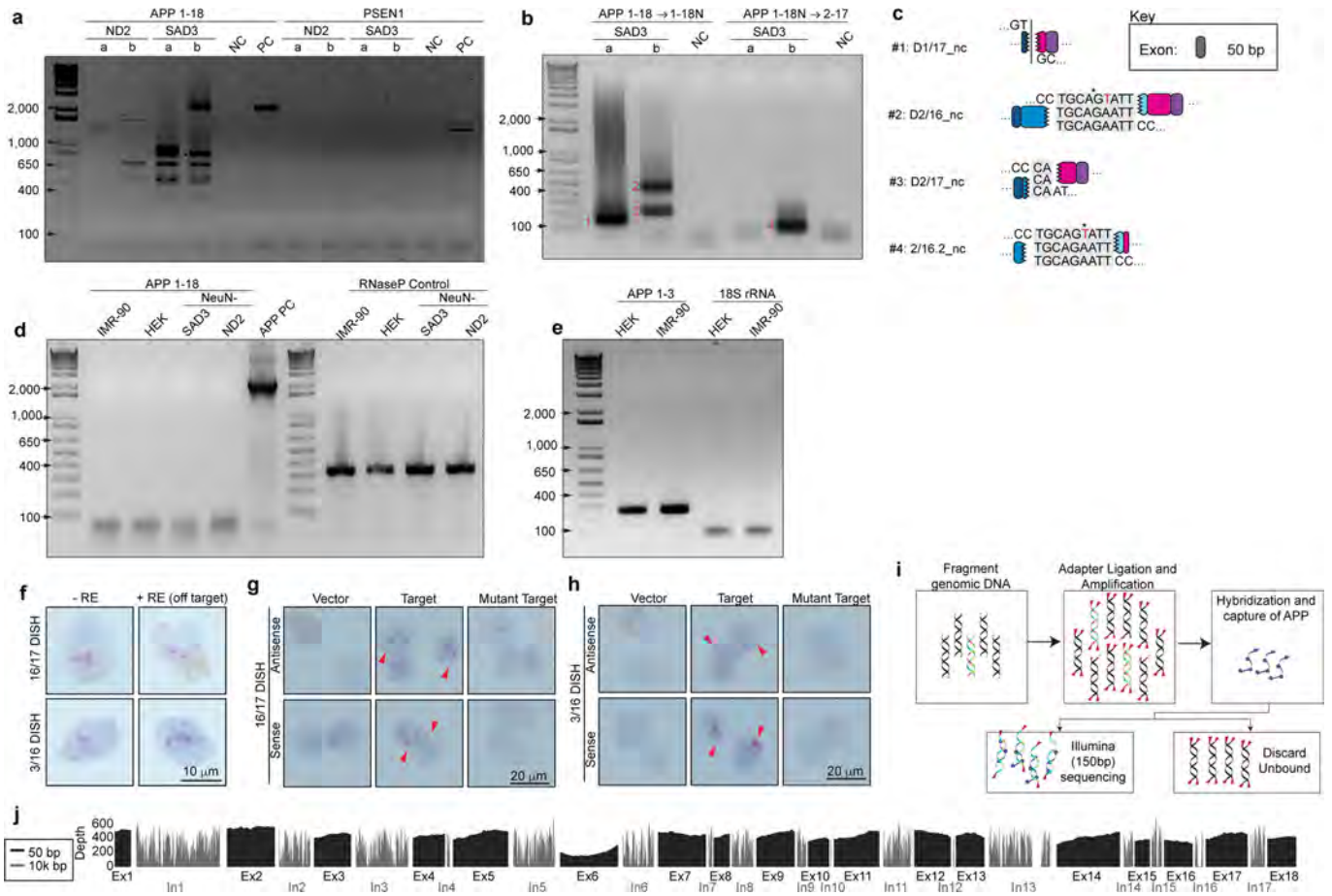
Data availability

Fastq files of SMRT sequences performed on a PacBio Sequel and Illumina sequences on NextSeq500 have been deposited in NCBI Sequence Read Archive (BioProject ID: PRJNA493258). The PacBio produced RNA-seq data sets from whole brain and temporal lobe supporting the findings of this study are available at https://www.pacb.com/blog/data-release-alzheimer-brain-isoform-sequencing-iso-seq-dataset, and from the authors upon reasonable request and with permission of PacBio, respectively. The source codes of the customized algorithms are available on GitHub (https://github.com/christine-liu/exonjunction and https://github.com/taolonglab/varccs).

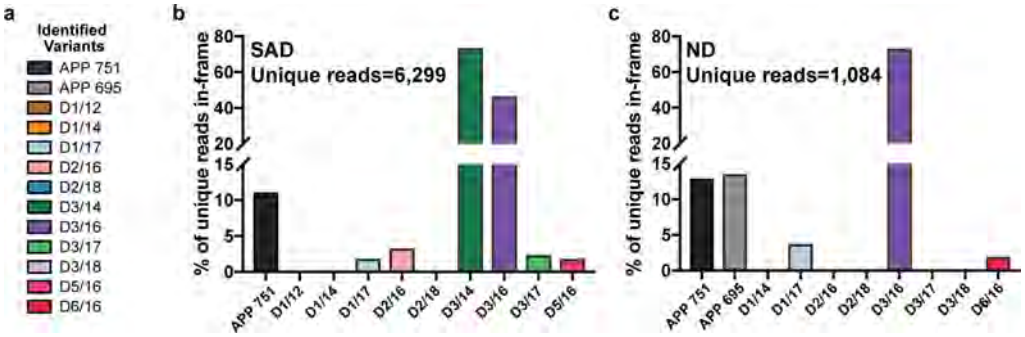
43. Hebert, P. D. N. et al. A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* **19**, 219 (2018).
44. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
45. Zhang, H., Susanto, T. T., Wan, Y. & Chen, S. L. Comprehensive mutagenesis of the fimS promoter regulatory switch reveals novel regulation of type 1 pili in uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **113**, 4182–4187 (2016).
46. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* **14**, 405 (2013).
47. Lovatt, A. et al. High throughput detection of retrovirus-associated reverse transcriptase using an improved fluorescent product enhanced reverse transcriptase assay and its comparison to conventional detection methods. *J. Virol. Methods* **82**, 185–200 (1999).
48. Ma, Y. K. & Khan, A. S. Evaluation of different RT enzyme standards for quantitation of retroviruses using the single-tube fluorescent product-enhanced reverse transcriptase assay. *J. Virol. Methods* **157**, 133–140 (2009).



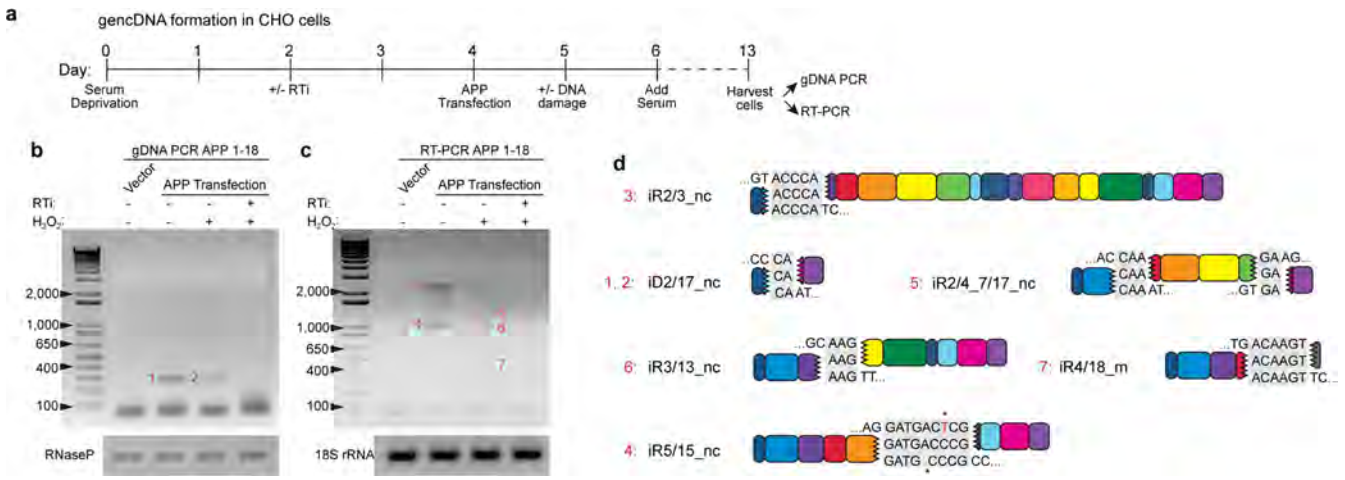
Extended Data Fig. 1 | RT–PCR on bulk and sorted nuclei, and RISH. a, RT–PCR products from bulk brain tissue samples from three individuals with SAD and three without. Canonical *APP* splice variants and non-*APP* products were identified. **b**, Representative gels showing the presence of canonical *APP* splice variants (red arrows, $n = 2$ independent experiments). **c**, No *APP* variants were identified in NeuN-negative nuclei from individuals with or without SAD. The *18S* rRNA control verified the presence of RNA. Novel *APP* RNA variants were identified from oligo-dT primed cDNA libraries from 50-cell populations of neuronal nuclei (**d**, $n = 3$ biological replicates) and brains from individuals with Alzheimer's disease (**e**, commercially produced PacBio cDNA libraries). **f**, $RISH_{3/16}$ signal from antisense probes showed cytoplasmic distribution of *APP* 3/16 RNA. Negative control sense probes and a probe targeting the bacterial gene *DapB* showed no signal. **g**, *PSEN1* RT–PCR on populations of 50 nuclei from the brains of three individuals with SAD and three without showed no *PSEN1* RNA variants. The positive control (PC) is amplified from RNA extracted from bulk brain tissue. 18S rRNA control verified the presence of RNA.



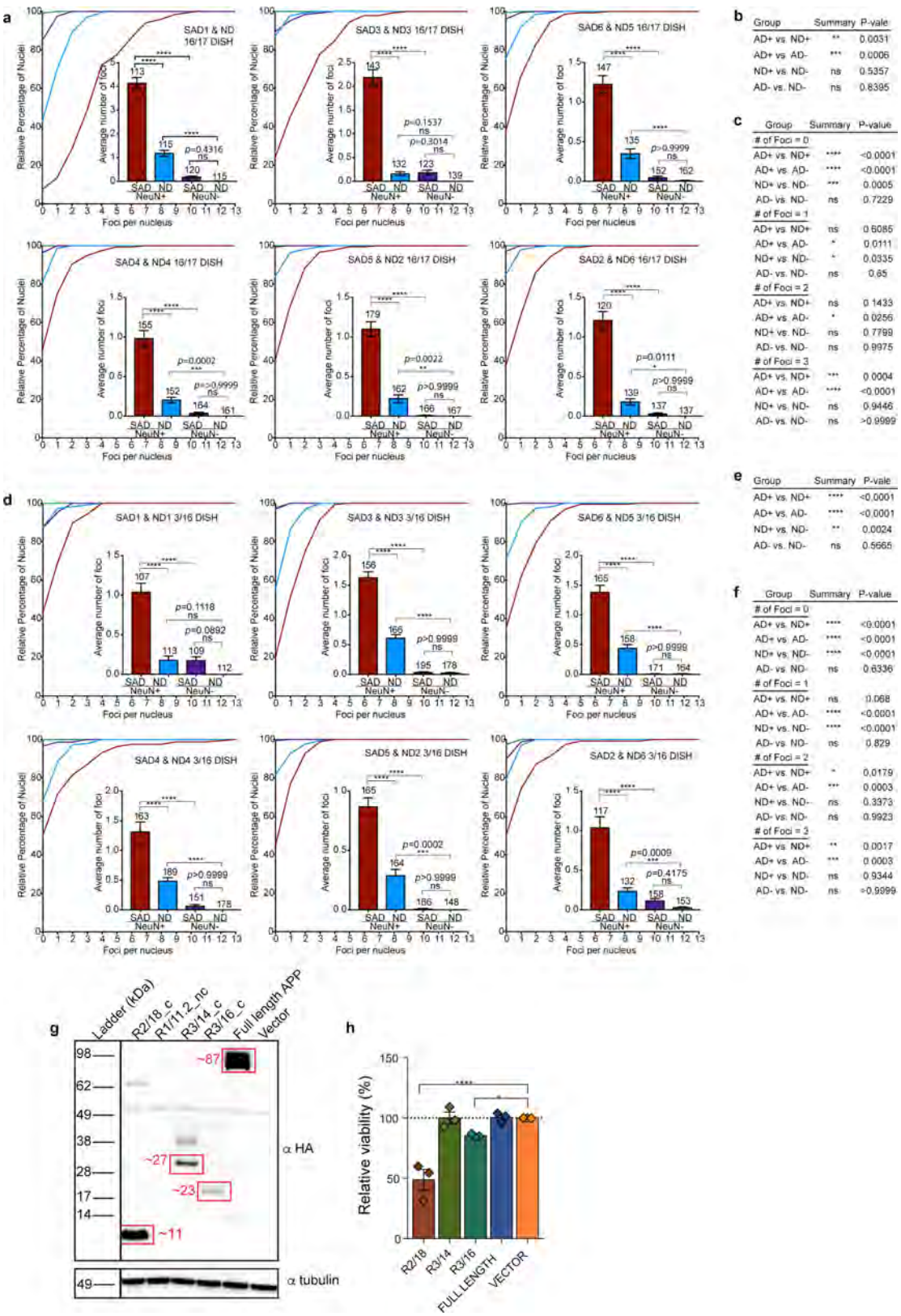
Extended Data Fig. 2 | *APP* gencDNA detection by genomic DNA PCR, DISH, and targeted genomic pull-down. a, Duplicate gel from Fig. 2b, with more sensitive thresholds to show the clear absence of *PSEN1* bands. **b**, Nested PCR was used with alternative *APP* primers (three total sets: *APP* 1–18, *APP* 1–18N, and *APP* 2–17). **c**, Cloning and Sanger sequencing of indicated bands (red numbers in **b**) revealed novel *APP* gencDNAs (see Fig. 1 for legend and nomenclature). **d**, *APP* 1–18 DNA PCR showed no products in non-neuronal cell types: IMR-90 (human lung fibroblast), HEK (human embryonic kidney) and non-neuronal (NeuN-negative) genomic brain DNA from individuals with and without SAD. RNaseP was used as a positive control. **e**, *APP* mRNA is expressed in HEK-293 and IMR-90 cells; *18S* rRNA used as a positive control. **f**, Digestion with the off-target restriction enzyme XbaI did not affect $DISH_{3/16}$ or $DISH_{16/17}$ signals. **g**, **h**, Synthetic DNA containing 16/17 (**g**) or 3/16 (**h**) target sequences (target), or wild-type human genomic *APP* sequences lacking IEJs and exon–exon junctions (mutant target) were introduced by retroviral transduction into NIH-3T3 cells. $DISH_{16/17}$ and $DISH_{3/16}$ signals from both sense and antisense probes were detected only in target infected cells. **i**, Schematic of Agilent SureSelect targeted DNA pull-down. **j**, Agilent SureSelect hybridization enrichment targeted the entire genomic locus of *APP* and showed unbiased sequencing depth across the full genomic locus. Exons and introns are shown on two scales.



Extended Data Fig. 3 | *APP* gencDNA reading frame analysis. a, Colour key for all gencDNAs with junctions identified by SMRT sequencing. **b**, **c**, Percentage of unique in-frame reads from brains of individual with SAD (**b**; 6,299 unique reads) or without SAD (**c**; 1,084 unique reads).

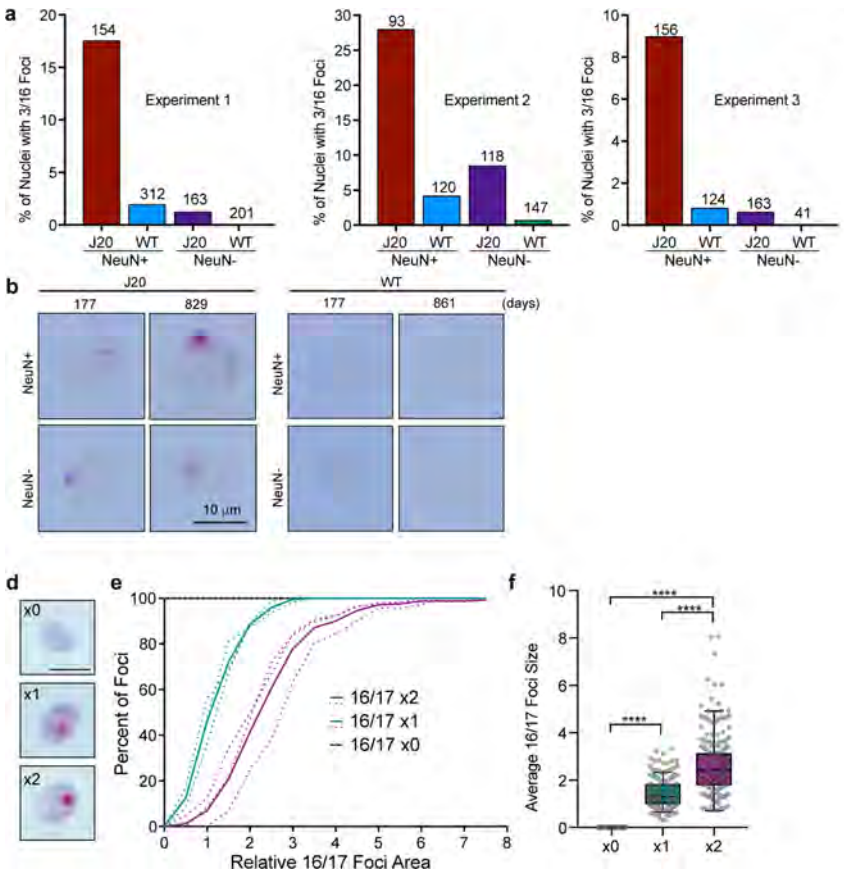


Extended Data Fig. 4 | *APP* gencDNA and RNA variant formation in CHO cells. a, Time line of CHO cell experiments modified from Fig. 4a. After transfection and gencDNA induction, serum was added and CHO cell cultures were passaged for 7 days. Cells were harvested, and DNA and RNA were extracted for analyses. **b**, **c**, PCR of genomic DNA (**b**; gDNA) and RT–PCR with APP 1 and 18 primers (**c**; $n = 2$ independent experiments). Note that APP plasmid is no longer detected (compare to Fig. 4b). DNA breaks during cell proliferation might contribute to variant formation in cells without DNA damage (no H_2O_2). Reverse transcriptase inhibitor (RTi, AZT + ABC) treatment prevents formation of *APP* RNA variants, indicating the dependence of RNA variants on gencDNAs. **d**, Induced *APP* variants with IEJs observed in **b**, **c**.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Data from six individual brains for each brain from individuals with or without SAD represented as averages in Fig. 5, and variant cytotoxicity. a, **d**, Nuclei sorted from cortices of six individuals with SAD and six without were analysed by $DISH_{16/17}$ (**a**) and $DISH_{3/16}$ (**d**). Cumulative frequency distribution plots and average numbers of foci per nucleus show statistical significance (non-parametric Kruskal–Wallis test with Dunn's correction for multiple comparisons) between all paired brain sets. Numbers above bars indicate number of nuclei analysed. NS, not significant. Error bars show s.e.m. **b**, **c**, **e**, **f**, Detailed *P* values for Fig. 5b (**b**), Fig. 5c (**c**), Fig. 5e (**e**) and Fig. 5f (**f**). **g**, *APP*-751, three coding and one non-coding *APP* variant in constructs containing haemagglutinin (HA) tags were transfected into HEK-293 cells. Cell lysates from all three coding variants and full-length *APP*-751 displayed protein products of the expected size by western blot. α-tubulin was used as a loading control. **h**, Three coding *APP* variants were transfected into SH-SY5Y cells individually and cell viability was measured by WST-1 seven days after transfection under serum-deprived conditions. Means of three independent experiments were analysed using ordinary one-way ANOVA with uncorrected Fisher's LSD for multiple comparisons ($*P = 0.0477$, $****P < 0.0001$).



Comparison	Group	Summary	P-value
Positive vs. negative	177+ vs. 177-	ns	>0.9999
	566+ vs. 566-	****	<0.0001
	661+ vs. 661-	***	0.0002
	728+ vs. 728-	****	<0.0001
	748+ vs. 748-	****	<0.0001
	829+ vs. 829-	****	<0.0001
Positive vs. Positive	177+ vs. 566+	****	<0.0001
	177+ vs. 661+	****	<0.0001
	177+ vs. 728+	****	<0.0001
	177+ vs. 748+	****	<0.0001
	177+ vs. 829+	****	<0.0001
	566+ vs. 661+	ns	>0.9999
	566+ vs. 728+	***	0.0007
	566+ vs. 748+	ns	0.2055
	566+ vs. 829+	**	0.0082
	661+ vs. 728+	****	<0.0001
	661+ vs. 748+	***	0.0009
	661+ vs. 829+	****	<0.0001
	728+ vs. 748+	ns	>0.9999
	728+ vs. 829+	ns	>0.9999
	748+ vs. 829+	ns	>0.9999
Negative vs. negative	177- vs. 566-	ns	>0.9999
	177- vs. 661-	ns	>0.9999
	177- vs. 728-	ns	0.5971
	177- vs. 748-	ns	0.0561
	177- vs. 829-	**	0.0026
	566- vs. 661-	ns	>0.9999
	566- vs. 728-	ns	>0.9999
	566- vs. 748-	ns	>0.9999
	566- vs. 829-	ns	>0.9999
	661- vs. 728-	ns	>0.9999
	661- vs. 748-	ns	0.8495
	661- vs. 829-	ns	0.1125
	728- vs. 748-	ns	>0.9999
	728- vs. 829-	ns	>0.9999
	748- vs. 829-	ns	>0.9999

Extended Data Fig. 6 | DISH$_{3/16}$ and DISH$_{16/17}$ data analyses. a, DISH$_{3/16}$ data from individual J20 and wild-type mouse cortices represented as an average in Fig. 5h; numbers above bars represent number of nuclei analysed. **b**, No DISH$_{16/17}$ signal was detected in wild-type mouse nuclei. **c**, Detailed statistical significance of DISH$_{16/17}$ signal across all mice in Fig. 5j (non-parametric Kruskal–Wallis with Dunn's multiple comparisons test). $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. NS, not significant. **d**–**f**, Synthetic DNA targets containing the exon 16/17 junction sequence were introduced by retroviral transduction into NIH-3T3 cells, and the target sequence (provirus) identified by DISH$_{16/17}$. A concatamer (×2) showed increased focus size, represented as a cumulative frequency distribution plot (**e**) and a box and whisker plot (**f**). Line, median; box, 75th–25th percentiles; whiskers, 90th–10th percentiles. Statistical significance was calculated using non-parametric Kruskal–Wallis test with Dunn's correction for multiple comparisons. $^{****}P < 0.0001$.

Extended Data Table 1 | Nine distinct experimental approaches supporting *APP* recombination

	Method	Tested Material	Unit Size of Sample	Reproducibility	Result
1	RT-PCR and Sanger sequencing	Nuclear RNA from human cerebral cortical neurons	50-nuclei	(1) Multiple brains (2) Multiple RT-PCRs (3) Multiple primers (4) Multiple investigators	Novel **APP RNA variants** with IEJs from multiple brains and experiments
2	RISH on IEJ 3/16	Human SAD tissue sections	Tissue sections	(1) Multiple sections (2) Multiple investigators	Cytoplasmic and mosaic **IEJ 3/16** signals on human brain tissue
3	Whole transcriptome SMRT sequencing	Human AD brain RNA prepared by PacBio	Bulk RNA	Multiple scientists using independent approaches identified IEJ variants	Novel **APP RNA variants** with IEJs
4	Targeted RNA SMRT sequencing	Human temporal lobe RNA, pulldown of APP related genes by PacBio	Bulk RNA	Multiple scientists using independent approaches identified IEJ variants	Novel **APP RNA variants** with IEJs
5	PCR and Sanger sequencing	Genomic DNA from SAD and non-diseased cerebral cortical neurons and non-neurons	DNA equivalent to 20-nuclei	(1) Multiple brains (2) Multiple PCRs (3) Multiple primers (4) Multiple polymerases (5) Multiple investigators	Conserved, 8 **gencDNAs** including the same IEJs in genomic DNA as found in RNAs
6	PCR and SMRT sequencing	Genomic DNA from SAD and non-diseased cerebral cortical neurons	Small nuclei populations (20-1000)	(1) Multiple brains (2) Multiple sequencing runs	At least 6,299 unique *APP* **gencDNA** variants from AD brain; 1,084 from non-diseased brain
7	Sequencing of *APP* genomic locus, pulled down by Agilent SureSelect	Genomic DNA from SAD cerebral cortical neurons	200 ng of DNA	Two complete sets of pull-down and sequencing	*APP* **gencDNAs**
		SAD and non-diseased cerebral cortical neuronal nuclei			Up to 13 spatially distinct *APP* **gencDNA** loci in single neuronal nucleus, enriched in AD and rare in non-neuronal nuclei
8	DISH on gencDNAs	J20 and WT cerebral cortical neuronal vs. non-neuronal nuclei	Single nuclei	(1) Multiple investigators (2) Multiple brains (3) Multiple sense probes	Increased ***APP* IEJ 3/16 gencDNA** signal in J20 neurons, but not non-neuronal nuclei
		J20 and WT cerebral cortical neuronal vs. non-neuronal nuclei			Age-related increases in ***APP* Ex 16/17 foci** diameter in J20 neuronal but not non-neuronal nuclei
9	APP751 over-expression in CHO cells with H_2O_2 treatment	CHO cells	Bulk DNA	(1) Multiple experiments (2) Multiple investigators	*APP* **gencDNA** with IEJs identified, reverse transcriptase activity and DNA breaks dependent

Evidence of *APP* recombination reported in this paper is summarized with experimental methods, materials, unit sizes of sample and reproducibility.

Extended Data Table 2 | Human postmortem brain information

Brain Name	Braak	Sex	PMI (Hours)	Age (years)
SAD-1	6	F	6	88
SAD-2	6	F	12	88
SAD-3	6	F	6	84
SAD-4	6	F	4	86
SAD-5	6	M	5	83
SAD-6	6	F	10	72
SAD-7	5	F	3.7	77
ND-1	1	M	U	87
ND-2	1	F	72	83
ND-3	U	M	U	83
ND-4	1	F	12	80
ND-3	1	F	18	93
ND-6	2	M	12	94

F, female; M, male; U, unknown; PMI, post mortem interval. All brains were from the pre-frontal cortex and obtained from the University of California San Diego Alzheimer's Disease Research Center and the University of California Irvine Institute for Mind Impairments and Neurological Disorders.

Extended Data Table 3 | *APP* variants information

Name	RNA PCR	DNA PCR	Coding or Non-coding	Start (bp)	Break Start	Break End	End	Sanger Sequence Homology	# of bp in homology	# of mis-matches
RT-PCR Identified Variants										
R1_11.1	Y	Y	Non-Coding	1	32	1431	2313	CACTGCTCTGCAGGC	15	3
R1_11.2	Y		Non-Coding	1	44	1456	2313	CGGC	4	0
R1_14	Y	Y	Non-Coding	1	46	1814	2313	AGCTC	5	1
R2_14	Y		Non-Coding	1	200	1749	2313	ACCAAGGA	8	0
R2_16	Y		Non-Coding	1	216	2015	2313	AT	2	0
R2_17	Y	Y	Non-Coding	1	64	2102	2313	CA	2	0
R2_18	Y	Y	Coding	1	211	2267	2313	GC	2	0
R3_14	Y	Y	Coding	1	267	1890	2313	AGCCAAC	7	0
R3_16	Y	Y	Coding	1	251	2008	2313	AA	2	0
R3_17	Y	Y	Non-Coding	1	314	2123	2313	GCAGTG	6	0
R6_17	Y		Coding	1	673	2079	2313	AGATGGGAGTGAAGACAAAG	20	0
R6_18	Y		Coding	1	740	2233	2313	GAGGA	5	0
DNA PCR Identified Variants										
D2_18		Y	Coding	1	120	2287	2310	N/A	N/A	N/A
D1_17		Y	Non-Coding	18	51	2159	2285	N/A	N/A	N/A
D2_16		Y	Non-Coding	19	209	2016	2285	TGCAGAATT	9	1
D2_17		Y	Non-Coding	18	64	2102	2285	CA	2	0
D2_16.2		Y	Non-Coding	157	209	2016	2095	TGCAGAATT	9	1
Commercially available PacBio RNA-Seq										
P3_9	Y	n/a	Non-coding	303	345	1093	+853	CCTAC	5	0
P6_12	Y	n/a	Non-coding	-41	724	1483	+853	AAGAAG	6	0
P6_18	Y	n/a	Non-coding	-111	2029	2274	+936	AATTCCGAC	9	0
DNA PCR on CHO cells: Induced Variants										
iD1_17	n/a	Y	Non-Coding	1	51	2159	2313	N/A	N/A	N/A
iD2_13	n/a	Y	Coding	1	170	1626	2313	N/A	N/A	N/A
iD4_15	n/a	Y	Missense	1	434	1920	2313	TGA	3	1
iD6_18	n/a	Y	Coding	1	705	2269	2313	T	1	0
iD2_17	n/a	Y	Non-Coding	1	64	2102	2313	CA	2	0
RNA PCR on CHO cells: Induced Variants										
iR2/3	Y	n/a	Non-Coding	1	64	319	2313	ACCCA	5	0
iR5/15	Y	n/a	Non-Coding	1	618	1921	2313	GATGACTCG	9	2
iR2/4_ 7/17	Y	n/a	Non-Coding	1	197/ 934	399/ 2077	2313	CAA/GA	3/2	0/0
iR3/13	Y	n/a	Non-Coding	1	318	1682	2313	AAG	3	0
iR4/18	Y	n/a	Missense	1	397	2285	2313	ACAAGT	6	0

Detailed information on identified *APP* RNA and DNA variants.

Extended Data Table 4 | DISH and RISH experiments and validation list

Junction	Target	Sample	Type	Probes	Figure Panel
16/17	DNA	Human nuclei +RNase	Exp	Sense	Fig. 2d,f,g,n; Fig. 5a,b,c
			Exp	Antisense	Fig. 2d,f,g
		Human nuclei +RNase + restriction enzyme (MluCI)	Neg	Sense	Fig. 2j,k
		Human nuclei +RNase +off-target restriction enzyme (XbaI)	Pos	Sense	ED Fig. 2f
		Synthetic target	Pos	Sense	ED Fig. 2g
			Pos	Antisense	ED Fig. 2g
		Synthetic mutant target	Neg	Sense	ED Fig. 2g
			Neg	Antisense	ED Fig. 2g
		Synthetic target concatamer	Pos	Sense	ED Fig. 6d-f
		WT mouse nuclei +RNase	Neg	Sense	Fig. 5i,j; ED Fig. 6b,c
		J20 mouse nuclei +RNase	Exp	Sense	Fig. 5i,j; ED Fig. 6b,c
3/16	DNA	Human nuclei +RNase	Exp	Sense	Fig. 2h,e,i,n; Fig. 5d,e,f
			Exp	Antisense	Fig. 2h,e,i
		Human nuclei +RNase + restriction enzyme (PSTI & MsII)	Neg	Sense	Fig. 2l,m
		Human nuclei +RNase +off-target restriction enzyme (XbaI)	Pos	Sense	ED Fig. 2f
		Synthetic target	Pos	Sense	ED Fig. 2h
			Pos	Antisense	ED Fig. 2h
		Synthetic mutant target	Neg	Sense	ED Fig. 2h
			Neg	Antisense	ED Fig. 2h
		WT mouse nuclei +RNase	Exp	Sense	Fig. 5g,h; ED Fig. 6a
		J20 mouse nuclei +RNase	Exp	Sense	Fig. 5g,h; ED Fig. 6a
	RNA	SAD tissue	Neg	Sense	ED Fig. 1f
			Exp	Antisense	ED Fig. 1f
			Neg	DapB	ED Fig. 1f
In2/Ex3	DNA	Human nuclei +RNase	Exp	Sense	Fig. 2n

In, intron; Ex, exon; Exp, experimental; Neg, negative control; Pos, positive control; ED, Extended Data. DNA and RNA in situ hybridization experiments, positive and negative controls are summarized.

nature research

Corresponding author(s): Jerold Chun

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	DNA gel and Western blot: BioRad Molecular Imager ChemiDoc XRS+ and Image Lab Software 4.0. Southern blot: Fujifilm FLA-5100 phosphorimager. Microscopy: Zeiss AX10 Imager. M2 microscope and ZEN2 software. PacBio CCS sequencing: PacBio Sequel sequencing platform. Sequencing data was processed from pre-CCS (circular consensus sequence) bam format to fastq using the publicly available circular consensus program in SMRTLink. Software analysis pipelines used BLAST 2.6.0, STAR 2.5.3a, and Picard 2.1.1 as described in the methods section. QPCR: BioRad CFX384 RealTime System and BioRad CFX Manager 3.1. Illumina sequencing: Illumina NextSeq 500. Sequences were aligned to the human reference genome (GRCh38) using STAR (version 2.5.3a). Duplicate reads were marked and removed using Picard (version 2.1.1).
Data analysis	ImageJ 1.50i and Graphpad Prism 7 were used for quantitative analysis of ISH. Appropriate software pipelines were used to classify and characterize sequencing data, these included SMRTLink 5.1.0.26412 and SMRTTools 5.1.0.26366 from Pacific Biosciences for analysis of SMRT sequencing results and STAR 2.5.3a and Picard 2.1.1 for Illumina sequencing. Reads were informatically analyzed using IGV 787, the UCSC Genome Browser (GRCh38/hg38), and the custom pipeline created as described in the methods section. Code will be uploaded to GitHub upon acceptance of the manuscript.

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 upon request. 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 (Fig. 2q, Extended Data Fig. 2i) and PacBio long read sequences (Fig. 3) used in the analysis will be made available on the NCBI Sequence Read Archive upon acceptance of the manuscript. Sanger sequences are provided in Supplementary Information.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-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, the experience of the authors, and comparison with similar published studies in order to ensure appropriate statistical power.
Data exclusions	No data was excluded from analysis.
Replication	All attempts at replication were successful. Replications of ISH and cell viability assays were all provided in Figures and Extended Data Figures. A summary of reproducibility for each experiment is shown in Extended Data Table 1.
Randomization	Samples were allocated randomly.
Blinding	All ISH counts were blinded for quantification and assessment. Automatic, unbiased quantification of foci number and size was performed via imageJ.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Fresh frozen human brain tissues were obtained from the University of California San Diego (UCSD) Alzheimer's Disease Research Center (ADRC) and the University of California Irvine (UCI) Institute for Mind Impairments and Neurological Disorders (MIND). Material limitations are present for post-mortem human brain analysis, where the reasons for restrictions included limited sample quantity with requests to individual brain banks granted on a case by case basis. Brain sample and donor data are provided in Extended Data Table 2.

Antibodies

Antibodies used	All antibodies used are listed (clone number, dilution, supplier, catalog number) FANS (cell sorting): 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) Western blot: Rat anti-HA antibody (3F10, 1:500, Roche,11867423001) Mouse anti-a-tubulin antibody (DM1A, 1:500, Sigma, T9026) HRP-conjugated goat anti-rat antibody (N/A, 1:10000, Cell Signaling, 7077S) HRP-conjugated goat anti-mouse antibody (N/A, 1:10000, Pierce, 1858413)
Validation	These antibodies are all published and validated by immunofluorescence staining (anti-NeuN, anti-a-tubulin, anti-rabbit), immunohistochemistry (anti-NeuN and anti-a-tubulin), immunoprecipitation (anti-HA) and Western blot (anti-NeuN, anti-HA, anti-a-tubulin, anti-rat and anti-mouse). Additional validation and peer-reviewed papers are available on the manufacturer's websites.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	NIH-3T3, CHO-K1, SH-SY5Y, IMR-90 and HEK-293 cells were all purchased from ATCC.
Authentication	Validated by ATCC.
Mycoplasma contamination	Cells from ATCC and all reagents used are verified Mycoplasma free.
Commonly misidentified lines (See ICLAC register)	HEK cells were used to show protein expression of our targets in mammalian cells no APP gencDNA as a negative control. For the indicated purpose, there is no concern regarding this cell line in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	J20 (strain: B6.Cg-Zbtb20Tg(PDGFB-APPSwInd)20Lms/2Mmjax, genetic background: C57BL/6) mice were purchased from The Jackson Laboratory. J20 transgenic mice and wild type controls used for experiments were obtained from crosses of mice hemizygous for the transgene with wild type mice from the colony. Sex and age (day) of mice used for experiments are listed : F177, M566, M661, M661, M728, F748, F829, and M861. All animals were cared under Sanford Burnham Prebys Medical Discovery Institute IACUC guidelines.
Wild animals	This study did not involve wild animals
Field-collected samples	This study did not involve wild animals