Recurrent patterns of widespread neuronal genomic damage shared by major neurodegenerative disorders

Zinan Zhou[1,2]†, Lovelace J. Luquette[3]†, Guanlan Dong[1,2,4]†, Junho Kim[5], Jayoung Ku[1,2], Kisong Kim[1,2], Mingyun Bae[1,2], Diane D. Shao[1,2,6], Bezawit Sahile[7], Michael B. Miller[1,2,8,9], August Yue Huang[1,2,9], William J. Nathan[10], Andre Nussenzweig[10], Peter J. Park[3]‡*, Clotilde Lagier-Tourenne[11]‡*, Eunjung Alice Lee[1,2,9]‡*, Christopher A. Walsh[1,2,9,12]‡*

Affiliations:

[1] Division of Genetics and Genomics, Manton Center for Orphan Disease Research, Boston Children’s Hospital; Boston, MA, USA.

[2] Department of Pediatrics, Harvard Medical School; Boston, MA, USA.

[3] Department of Biomedical Informatics, Harvard Medical School; Boston, MA, USA.

[4] Bioinformatics and Integrative Genomics Program, Harvard Medical School; Boston, MA, USA.

[5] Department of Biological Sciences, Sungkyunkwan University; Suwon, South Korea.

[6] Department of Neurology, Boston Children’s Hospital; Boston, MA, USA.

[7] Program in Neuroscience, Harvard Medical School; Boston, MA, USA.

[8] Division of Neuropathology, Department of Pathology, Brigham and Women’s Hospital, Harvard Medical School; Boston, MA, USA.

[9] Broad Institute of MIT and Harvard; Cambridge, MA, USA.

[10] Laboratory of Genome Integrity, National Cancer Institute, NIH, Bethesda, MD, USA

[11] Department of Neurology, Sean M. Healey & AMG Center for ALS, Massachusetts General Hospital, Harvard Medical School; Boston, MA, USA.

[12] Howard Hughes Medical Institute; Boston, MA, USA.

*Corresponding author. Email: peter_park@hms.harvard.edu (P.J.P); CLAGIER-TOURENNE@mgh.harvard.edu (C.L.-T.); ealee@childrens.harvard.edu (E.A.L.) christopher.walsh@childrens.harvard.edu (C.A.W);

† These authors contributed equally.

‡ These authors jointly supervised this work.

Abstract

Amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD), and Alzheimer's disease (AD) are common neurodegenerative disorders for which the mechanisms driving neuronal death remain unclear. Single-cell whole-genome sequencing of 429 neurons from three *C9ORF72* ALS, six *C9ORF72* FTD, seven AD, and twenty-three neurotypical control brains revealed significantly increased burdens in somatic single nucleotide variant (sSNV) and insertion/deletion (sIndel) in all three disease conditions. Mutational signature analysis identified a disease-associated sSNV signature suggestive of oxidative damage and an sIndel process, affecting 28% of ALS, 79% of FTD, and 65% of AD neurons but only 5% of control neurons (diseased vs. control: OR=31.20, $p = 2.35\times10^{-10}$). Disease-associated sIndels were primarily two-basepair deletions resembling signature ID4, which was previously linked to topoisomerase 1 (TOP1)-mediated mutagenesis. Duplex sequencing confirmed the presence of sIndels and identified similar single-strand events as potential precursor lesions. TOP1-associated sIndel mutagenesis and resulting genome instability may thus represent a common mechanism of neurodegeneration.

Introduction

Amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD) and Alzheimer's disease (AD)(*1, 2*) are common neurodegenerative diseases with diverse pathologies and genetic underpinnings. ALS and FTD are closely related, with clinical, pathological and genetic overlap, marked by the depletion of nuclear TDP-43 and accumulation of cytoplasmic TDP-43 inclusions in neurons. Mutations in many genes such as *C9ORF72*, *TARDBP (TDP-43)*, *FUS, TBK1, VCP,* and *SQSTM1* contribute to both diseases(*3*), with *C9ORF72* repeat expansion being the most common genetic cause of both familial and sporadic forms of ALS and FTD(*4, 5*). In contrast, AD pathology is characterized by accumulation of amyloid-β protein and phosphorylated tau, and shows distinct underlying genetic risks(*6*).

Genomic instability, which leads to the accumulation of DNA damage and mutations, has been implicated in all of these conditions, but its exact role remains elusive. In ALS, FTD and AD, increased DNA damage due to high levels of oxidative stress, R-loops and DNA strand breaks has been reported(*7-14*). Several ALS/FTD-related genes are involved in DNA damage response, such as *SOD1*, *FUS*, *TDP-43*, *C9ORF72*, *NEK1*, *SQSTM1*, *SETX,* and *VCP*(*9, 15-23*). *C9ORF72* repeat expansions have recently been shown to induce chromosomal fragility and DNA damage(*24*). Neurons are particularly vulnerable to DNA damage due to their high transcriptional activity, large energy demand and associated oxidative stress. Recent studies have shown that somatic single nucleotide variants (sSNVs) accumulate in aging neurons, with further increases observed in AD, Cockayne syndrome, and Xeroderma pigmentosum(*13, 14, 25*).

In the present study, we profiled sSNVs and somatic insertions and deletions (sIndels) in neurons with and without depletion of nuclear TDP-43 from the premotor cortex of three *C9ORF72* ALS and the prefrontal cortex of six *C9ORF72* FTD brains using single-cell whole-genome sequencing (scWGS). We then compared the results to those from AD neurons(*14*), in which sIndels were not previously analyzed, and neurons from neurotypical controls. Our analysis revealed significantly elevated levels of sSNVs and sIndels in neurons from *C9ORF72* ALS, *C9ORF72* FTD and AD brains despite their diverse genetic basis and pathologies. Importantly, our findings indicate that neurons in all three disease conditions often exhibit an excessive number of sIndels—sometimes over 1000, equivalent to hundreds of years of age-related sIndel accumulation—and share a mutational pattern characterized by two-basepair (2-bp) deletions. Mutational signature analysis of these sIndels associates the mutagenic process to topoisomerase 1 (TOP1)-mediated mutagenesis. Our results suggest that increased levels of DNA damage and accelerated accumulation of somatic mutations likely contribute to the pathogenesis of *C9ORF72* ALS, *C9ORF72* FTD and AD.

Results

Isolation of neuronal nuclei with TDP-43 pathology from *C9ORF72* ALS and FTD brains achieves high purity

To isolate neuronal nuclei with either normal or depleted nuclear TDP-43 (TDP-43+ and TDP-43-) from postmortem *C9ORF72* ALS and FTD brains, we performed fluorescence-activated nuclear sorting (FANS) with co-staining of NeuN and TDP-43 antibodies(*26*) (Fig. 1A). TDP-43- neurons were only found in *C9ORF72* ALS and FTD brains and were absent in age-matched controls (Fig. 1B). Analysis of single-nucleus RNA sequencing (snRNA-seq) data from isolated nuclei revealed that 96.4% of TDP-43+ and 95.4% of TDP-43- neuronal nuclei were indeed neuronal (Fig. 1C and fig. S1A). Furthermore, we examined cryptic exons caused by loss of TDP-43 function during RNA splicing in the snRNA-seq data(*27-29*). Transcripts containing known cryptic exons in the *RAP1GAP*, *STMN2*, *ATP8A2* and *KALRN* genes were detected in TDP-43- neurons but absent or at very low levels in TDP-43+ neurons (Fig. 1, D and E and fig. S1, B and C), confirming the presence of TDP-43 disease pathology in isolated TDP-43- neurons.

Both TDP-43+ and TDP-43- neurons encompassed all major clusters of excitatory and inhibitory neurons (Fig. 1C), though TDP-43- neurons were notably depleted in inhibitory neurons (Fig. 1C and fig. S2A), confirmed by the reduced expressions of inhibitory neuron markers in TDP-43- neurons compared to TDP-43+ neurons in previously generated bulk RNA-seq data from *C9ORF72* FTD brains (fig. S2B)(*26*). These results are consistent with recent studies demonstrating that inhibitory neurons rarely exhibit cytoplasmic TDP-43 inclusions(*30*). However, it remains unclear whether the reduced TDP-43 pathology in inhibitory neurons reflects resistance to TDP-43 pathology or susceptibility to TDP-43 pathology given the observed loss of inhibitory neurons in ALS(*31-33*).

Somatic mutation burdens are increased in diseased neurons

We performed primary template-directed amplification (PTA)-based single-cell whole-genome amplification (scWGA) and sequencing (scWGS) on neurons isolated from the premotor cortex—the most disease relevant region available—of three *C9ORF72* ALS brains and the prefrontal cortex of six *C9ORF72* FTD brains (three TDP-43+ and three TDP-43- neurons per brain; table S1 and S2), and identified sSNVs and sIndels using Single Cell ANalysis 2 (SCAN2)(*25*), our recently developed single-cell genotyper (Fig. 1A). Somatic mutations in these neurons were compared with those in 24 newly sequenced premotor cortex neurons and 56 previously sequenced prefrontal cortex neurons from neurotypical controls across various ages(*25, 34*), as well as 29 previously sequenced prefrontal cortex neurons from AD brains, for which sIndels were not previously studied (Fig. 1A and tables S1 and S2)(*14*). Cells that did not pass quality control (QC) criteria, such as amplification uniformity, were excluded from downstream analyses (table S2, Methods). SCAN2 recovered approximately 38% of sSNVs and 26% of sIndels, producing a somatic mutation catalog with >82,000 sSNVs and >12,000 sIndels with an aggregate error rate of <10% (fig. S3).

We first compared the burdens of sSNVs and sIndels between neurons from the prefrontal cortex and premotor cortex in neurotypical controls aged >50 years. Neurons from both regions had similar burdens of sSNVs and sIndels (Fig. 2A) and were therefore combined for subsequent analyses. We then compared somatic mutations in neurons from *C9ORF72* ALS, *C9ORF72* FTD and AD brains to neurons from neurotypical brains. Burdens of both sSNVs and sIndels in neurons from all three neurodegenerative conditions were significantly increased compared to those from neurotypical controls after adjusting for age, with sIndels showing the greatest increases (Fig. 2, B and C). Individual neurons from diseased brains exhibited variability in sSNV and sIndel burdens, even within the same brain (fig. S4). Increases in somatic mutations in *C9ORF72* ALS neurons were relatively modest compared to *C9ORF72* FTD and AD neurons, perhaps due to the selective involvement and rapid disappearance of motor neurons in ALS, and the involvement of a broader range of neuronal types in FTD and AD(*35*). Together, our results show that accumulation of sSNVs and sIndels is accelerated in *C9ORF72* ALS, *C9ORF72* FTD and AD neurons. Unexpectedly, TDP-43- neurons from *C9ORF72* ALS and FTD brains did not exhibit higher burdens of sSNVs and sIndels compared to TDP-43+ neurons (Fig. 2D), despite TDP-43's involvement in DNA damage repair of double-strand breaks(*9*). This observation suggests that the roles of *C9ORF72* repeat expansion and depletion of nuclear TDP-43 in causing DNA damage are not synergistic.

Diseased neurons exhibit increased 2-bp deletions with a recurrent pattern

Strikingly, neurons from all three disease conditions exhibited greatly increased rates of 2-bp deletions compared to controls (Fig. 3, A and B). To investigate potential mutagenic mechanisms, we performed de novo mutational signature extraction, a technique that can identify underlying processes based on the types and frequencies of mutations (*36*). D*e novo* mutational signature extraction was performed on neurons from the three neurodegenerative conditions, along with neurons and glial cells from neurotypical controls amplified using PTA. To increase power to recover relevant mutational signatures, we also included 81 neurons from AD, as well as 190 neurons and 40 glial cells from neurotypical controls previously amplified using multiple displacement amplification (MDA), an earlier whole genome amplification (WGA) method(*37*). This analysis of sSNVs resulted in six *de novo* SNV signatures (SBS-A to SBS-F, fig. S5A) of which only SBS-B was significantly elevated in PTA neurons from all three neurodegenerative conditions compared to control PTA neurons (fig. S5, B and C). To explore the mutagenic processes associated with the *de novo* signatures, we compared them to existing mutational signatures in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. SBS-B closely resembles the COSMIC SBS30 signature (cosine similarity 0.89), which is associated with inactivating mutations in *NTHL1,* an enzyme involved in base excision repair of oxidative DNA damage(*38*). The elevation of SBS-B sSNVs in neurons from all three neurodegenerative conditions may therefore indicate increased levels of oxidative DNA damage, consistent with previous studies revealing increased oxidative stress in these diseases(*7, 10, 12*). SBS-A and SBS-D are similar to COSMIC SBS5 and SBS1, respectively, both of which are “clock-like” signatures. While SBS1 is cell-division dependent and thus limited in postmitotic neurons, SBS5 accumulates universally during aging (fig. S5C). The mechanisms underlying SBS-C and SBS-E (which show no increase in diseased neurons) and SBS-F (which is increased only in AD neurons) are unclear.

De novo mutational signature extraction of sIndels identified two distinct signatures, ID-A and ID-B (Fig. 3C). ID-A burdens were significantly increased in PTA neurons from the three neurodegenerative conditions compared to control PTA neurons (Fig. 3D and fig. S5D), while burdens of ID-B showed little difference across PTA neurons from all conditions. ID-A is dominated by 2-bp deletions and resembles the COSMIC signature ID4 (cosine similarity 0.78) (Fig. 3C), which has recently been linked to a deficiency in RNase H2, an enzyme responsible for removing ribonucleotides erroneously incorporated into the genome. In the absence of functional RNase H2, ribonucleotides are alternatively removed through the TOP1-mediated repair pathway(*39*). Although ID4 is a feature of normal aging in neurotypical neurons—even more so than the canonical "clock-like" indel signatures ID5 and ID8 identified in cancers(*25, 34, 40*)—the excessive ID-A burden we observe in all three neurodegenerative conditions suggests a dramatic acceleration of this age-associated process. Indeed, substantially all of the excess sIndel burdens in diseased neurons can be attributed to ID-A. We further constructed a classifier (Methods) to determine which cells exhibited ID-A burdens in line with typical neuronal aging (ID-A normal) or excessive ID-A burden (ID-A high). Upon stratifying our cells by ID-A status (Fig. 3E), we found significantly higher fractions of ID-A high neurons in every neurodegenerative condition (ALS: OR = 7.01 [1.00–81.91], $p = 0.025$; FTD: OR = 66.38 [12.60–688.58], $p = 2.10 \times 10^{-11}$; AD: OR = 33.23 [6.27–346.35], $p = 1.45 \times 10^{-7}$; brackets indicate 95% confidence intervals, all comparisons vs. control neurons) while only two out of 80 control neurons and none of the 66 control oligodendrocytes were classified as ID-A high. Further, 13 of the 16 diseased brains contained at least one ID-A-high neuron, and the mutational spectra of ID-A-high and -normal neurons were remarkably similar across conditions (Fig. 3F), supporting the notion that ID-A pathology is likely a universal feature of these diverse neurodegenerative conditions rather than an isolated occurrence.

ID-A deletions exhibited features of TOP1-mediated ribonucleotide excision repair

To further investigate the association between ID-A sIndels in diseased neurons and TOP1-mediated ribonucleotide excision repair (RER), we examined features of the excessive 2-bp deletions. Mammalian TOP1 exhibits a preference for thymine bases and cleaves at the 3' phosphodiester bond following the thymine nucleotide(*41*), primarily causing deletions at a TNT motif, with the most common deletions being CT dinucleotides (Fig. 4A)(*39*). Consistent with this observation, the 2-bp deletions in neurons from neurodegenerative conditions were enriched in CT deletions and occurred most frequently within TNT motifs (Fig. 4, B and C). The TOP1-mediated RER process involves multiples steps in which a transient protein complex, TOP1 cleavage complex (TOP1cc), is formed through TOP1-DNA-protein crosslinks (Fig. 4A). Consistent with this, flow cytometry revealed higher fractions of neurons with increased TOP1cc activity in several *C9ORF72* FTD and AD brains compared to control brains (fig. S6A), directly suggesting formation of TOP1 covalent links to DNA.

TOP1-mediated ID4 indels can arise during both replication and transcription in mitotic cells(*42*), and are enriched at transcribed regions in cells with RNase H2 deficiency(*39*). In postmitotic neurons, ID4 indels are also strongly associated with transcription(*34*), as these cells lack cell division. To explore the association between sIndels and transcription, we first split all sIndels into two categories: (1) the 2-bp sIndel types characteristic of ID-A (Fig. 3C) and (2) all other sIndels. We then compared the burdens of each sIndel category to local gene expression levels measured in excitatory neurons from previously published snRNA-seq data(*34*). In neurotypical control neurons, both sIndel categories were strongly associated with expression levels (Fig. 4D). In the neurodegenerative conditions (which contain higher amounts of ID-A-like sIndels), both categories remained positively associated with expression, though the association was weaker for ID-A-like sIndels (Fig. 4D). This association with transcriptional activity was replicated by comparing each sIndel category against local levels of open chromatin (measured by single-nucleus ATAC-seq), active and inactive histone marks(*43*), chromatin states(*44*), and promoters and enhancers active in neurons(*45*) (fig. S7, A and B). Taken together, these characteristics of TOP1-related indel mutagenesis suggest that ID-A-like sIndels are likely TOP1-mediated in these three neurodegenerative conditions.

Single-strand deletions and breaks are prevalent in neurodegenerative conditions

During TOP1-mediated RER, the removal of incorporated ribonucleotides creates single-strand breaks (Fig. 4A). To assess these single-strand breaks, denatured genomic DNA (gDNA) extracted from prefrontal cortex of *C9ORF72* FTD, AD, and control brains was analyzed using agarose gel electrophoresis (Methods). The levels of fragmented single-strand gDNA were increased in *C9ORF72* FTD and AD brains, correlating with the levels of ID-A sIndels in neurons (Fig. 4, E and F). During RER, the subsequent re-ligation of single-strand breaks creates an intermediate DNA product containing a 2-bp single-strand deletion. To assess the presence of single-strand deletions, we performed multiplexed end-tagging amplification of complementary strands (META-CS)(*46*), a type of duplex sequencing, which distinguishes the Watson and Crick strands of DNA molecules. We examined samples of 50 neurons isolated from each of three *C9ORF72* FTD brains with high levels of ID-A sIndels and three age-matched control brains. The duplex sequencing data confirmed the disease-specific enrichment of ID-A-like, double-strand 2-bp deletions, and further revealed a larger proportion of ID-A-like single-strand lesions in *C9ORF72* FTD brains (Fig. 4G). The proportions of ID-A deletions were significantly higher in *C9ORF72* FTD brains compared to control brains for both double- and single-strand events, where single-strand deletions showed a more prominent increase (Fig. 4H). This suggests that the increased levels of TOP1-mediated RER in neurons from neurodegenerative conditions generates frequent single-strand breaks and deletions, some of which become fixed as double-strand indels. Together, these findings indicate that increased TOP1-mediated RER is a common mechanism underlying widespread, catastrophic DNA damage of all three neurodegenerative conditions.

Deficiency of RNase H2 and ribonucleotide reductase in diseased neurons

During DNA repair, DNA polymerases must incorporate dNTPs and discriminate against rNTPs. This selectivity relies on polymerase properties, as well as the balance of dNTP and rNTP pools, which is regulated by ribonucleotide reductases (RNRs)(*47*): when the dNTP:rNTP ratio is low, rNTP incorporation increases, threatening genome stability. Embedded rNTPs are removed by RNase H2, but in its absence, TOP1 serves as an alternative repair pathway (Fig. 5A). We found that genes encoding RNRs and RNase H2 subunits were downregulated across various neuronal subtypes in *C9ORF72* FTD brains compared to controls (fig. S8A). A similar trend of reduced expression was also observed in previously published snRNA-seq data from AD neurons (fig. S8A), although most of these changes were not statistically significant, likely due to the low abundance of these transcripts in neurons, as well as differences in tissue lysis and nuclear isolation protocols used in this study. However, qPCR on neurons from several *C9ORF72* FTD and AD brains with high levels of ID-A sIndels and control brains confirmed reduced expression of *RNASEH2B* and *RRM2B* in *C9ORF72* FTD and AD neurons compared to control neurons (Fig. 5B and fig. S8B).

In addition to insufficient RNRs, increased DNA damage repair is also known to deplete cellular dNTP pools, leading to low dNTP:rNTP ratios and increased ribonucleotide incorporation into the genome. DNA polymerases β and λ, which participate in base excision repair of oxidative DNA damage(*48*), are particularly prone to erroneously incorporating ribonucleotides. Consistent with this, we observed a positive correlation between ID-A and SBS-B—which resembles COSMIC SBS30, a signature associated with oxidative DNA damage—burdens in diseased neurons (Fig. 5C and fig. S8C), suggesting that those neurons that accumulate higher levels of DNA damage due to oxidized nucleotides are the same neurons that show catastrophic increases in TOP1-mediated mutagenesis, possibly reflecting low dNTP:rNTP ratios due to increased DNA repair activity.

Discussion

Our results revealed an accelerated accumulation of both sSNVs and sIndels in neurons from *C9ORF72* ALS, *C9ORF72* FTD and AD brains. While sIndels accumulate at a much slower rate than sSNVs in normal neurons—reaching approximately 200 to 300 sIndels in elderly individuals(*25*)— this process is greatly accelerated in neurons affected by these three neurodegenerative conditions, with some neurons exhibiting several thousand sIndels (Fig. 2B), equivalent to hundreds of years of age-related sIndel accumulation. Intriguingly, the excessive sIndels in neurons from all three diseases exhibited a single, shared mutational signature resembling the COSMIC ID4 signature (Fig. 3, C and F). COSMIC ID4 has been linked to RNase H2 deficiency and alternative TOP1-mediated RER(*39*) and is also the most prevalent aging-associated indel signature in normal neurons, as demonstrated recently(*25*). Furthermore, we observed several other characteristics of TOP1-mediated indels, including the local TNT motif, the specific dinucleotide most commonly deleted (CT), the association with local gene expression level, and other expected side effects of TOP1 activity, such as a preponderance of single-strand breaks and precursor single-strand deletions. Taken together, this evidence strongly suggests a role for TOP1 in sIndel mutagenesis in neurodegenerative conditions of diverse phenotypes and genetic origin.

TOP1 is a DNA topoisomerase that facilitates transcription, DNA replication, and DNA repair by alleviating topological tensions in the DNA(*49*). During this process, TOP1 forms a transient cleavage complex (TOP1cc), in which it covalently binds to DNA and introduces a single-strand nick that is subsequently re-ligated. In neurons, TOP1 has been implicated in the transcriptional regulation of long neuronal genes essential for synaptic function and plasticity(*50, 51*), and depletion of TOP1 in neurons has been linked to early-onset neurodegeneration and autism(*50, 52*). In the absence of RNase H2, TOP1 compensates by facilitating the removal of misincorporated ribonucleotides from the genome. Single-strand breaks induced by TOP1 can produce intermediate DNA products with single-strand 2-bp deletions, which were detected by our duplex sequencing approach (Fig. 4, G and H). In some cases, re-ligation of single-strand breaks can be hindered by multiple endogenous and environmental factors, such as local oxidative DNA damage, which traps TOP1cc(*49, 53*), resulting in persistent DNA damage. Consistent with the increased oxidative stress observed in neurons from the three neurodegenerative conditions (fig. S5, A and B, SBS-B), the increased levels of single-strand breaks in *C9ORF72* FTD and AD brains is likely a result of TOP1cc trapping by local oxidative DNA damage (Fig. 4, E and F). The persistent, widespread DNA damage may ultimately lead to neuronal cell death given that it would likely create widespread transcriptional abnormalities.

Mutations in genes encoding the three subunits of RNase H2 are linked to Aicardi-Goutières syndrome(*54*), an early onset disease characterized by chronic neuroinflammation, progressive neurological decline, neurodegeneration, and movement disorders. The RNase H2 deficiency in Aicardi-Goutières syndrome activates the stimulator of interferon genes (STING) pathway and triggers the innate immune response. RNase H2 deficiency can activate the STING pathway through both the canonical cyclic GMP-AMP synthase (cGAS) pathway—upon sensing cytoplasmic DNA—and the non-canonical NF-κB pathway in response to DNA damage(*55*). Activation of the STING pathway has been demonstrated in ALS, FTD and AD brains(*56-59*). Together with the recent finding of cGAS and NF-κB pathway activation in neurons from familial and sporadic ALS cases(*59*), our results suggest a link between the pathogenesis of Aicardi-Goutières syndrome and late-onset neurodegenerative diseases. Furthermore, our data suggest that modulation of TOP1 and RER activity may represent new potential therapeutic targets in a broad spectrum of neurodegenerative conditions.

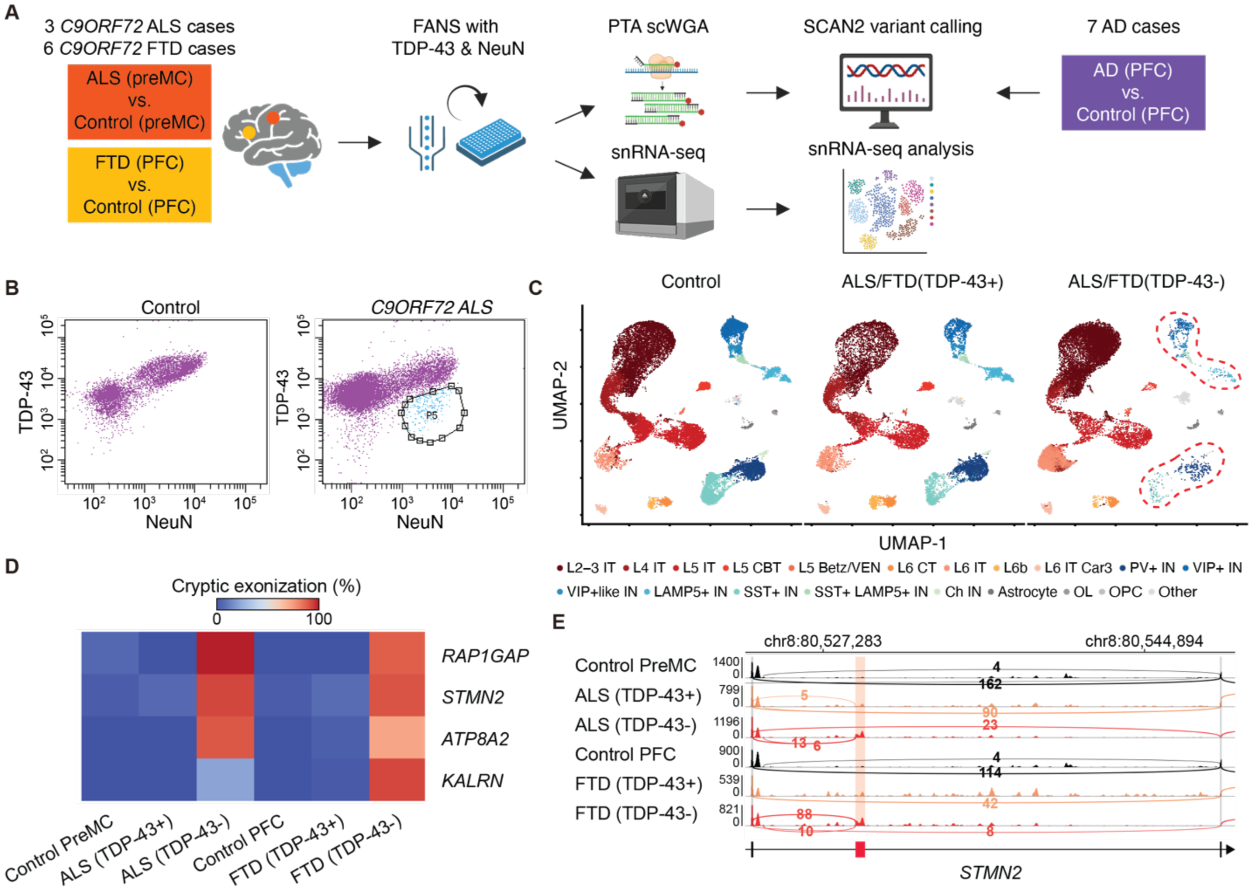

Fig. 1. Experimental strategy and characterization of isolated neuronal nuclei from *C9ORF72* ALS and FTD brains. (**A**) Overview of the experimental design. Neuronal nuclei with and without nuclear TDP-43 depletion were isolated from 3 *C9ORF72* ALS and 6 *C9ORF72* FTD brains, then subjected to genome sequencing and analysis. These were compared with neuronal nuclei from neurotypical brains, including newly generated data and previously published datasets. Previously generated scWGS data from AD brains were included. (**B**) Representative FACS plots showing neuronal nuclei from a *C9ORF72* ALS brain and a neurotypical control brain co-stained with NeuN and TDP-43 antibodies. The P5 gate (right, circle) indicates neuronal nuclei with depletion of nuclear TDP-43. (**C**) UMAP clustering of snRNA-seq data from neuronal nuclei, colored by major cell types. TDP-43- neuronal nuclei come exclusively from *C9ORF72* ALS and FTD brains, while TDP-43+ neuronal nuclei come from both diseased and neurotypical brains; control indicates neurotypical brains. IT: intratelencephalic neurons. CBT: corticobulbar tract neurons. VEN: Von Economo neurons. CT: corticothalamic neurons. PV: parvalbumin. IN: inhibitory neurons. Ch IN: cholinergic inhibitory neurons. OL: oligodendrocytes. OPC: oligodendrocyte precursor cells. Red dashed circles

indicate reduced inhibitory neurons under TDP-43- condition (**D**) Heatmap of cryptic exonization showing the proportion of transcripts with cryptic exons in neurons across different conditions. (**E**) IGV Genome browser tracks of *STMN2* showing snRNA-seq read coverage and splice junctions. TDP-43- neurons (red) contain a marked increase in cryptic exon inclusion (highlighted region), which is absent or low in TDP-43+ (orange) and control (black) neurons. Numbers next to splice junctions are read counts supporting the junction. PFC: prefrontal cortex. preMC: premotor cortex.

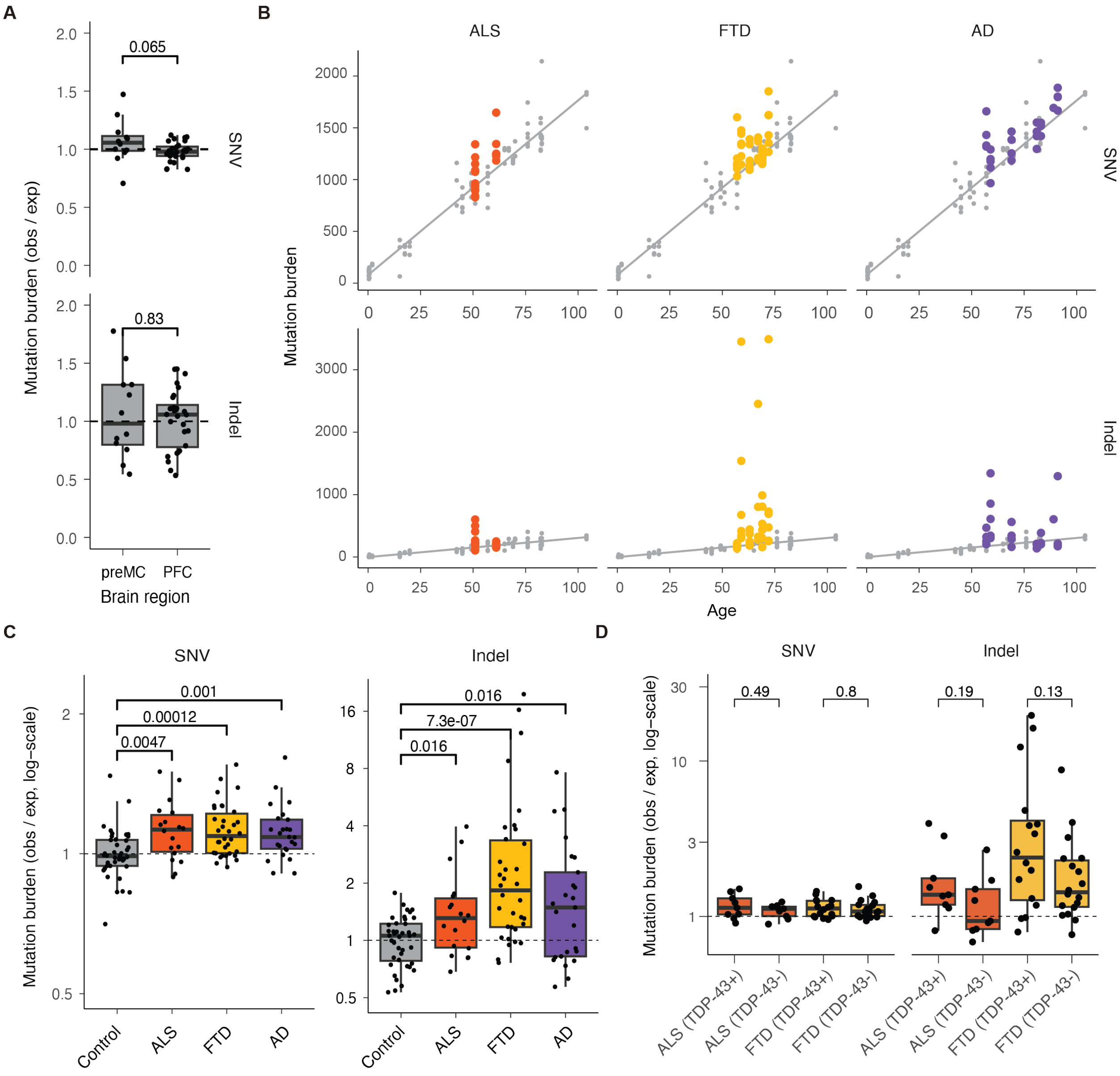

Fig. 2. Increased somatic mutation burdens in neurons from *C9ORF72* ALS, *C9ORF72* FTD and AD. (**A**) PFC (prefrontal cortex) and preMC (premotor cortex) neurons from neurotypical brains show comparable burdens of sSNVs and sIndels. Each point represents the burden of one neuron, compared against the expected burden for its age (Methods). (**B**) Extrapolated genome-wide sSNV and sIndel burdens for neurons from *C9ORF72* ALS, *C9ORF72* FTD and AD brains compared to those for neurons from neurotypical brains as a function of age. Identical control neuron points and regression (mixed-effects linear regression, Methods) are shown in gray in each panel for comparison. (**C**) Mutation burdens corrected for age (Methods); preMC and PFC neurons are combined to form the control set. (**D**) TDP-43+ and

TDP-43- neurons from *C9ORF72* ALS and FTD brains show similar levels of sSNVs and sIndels. All P-values in this figure represent Wilcoxon rank-sum tests.

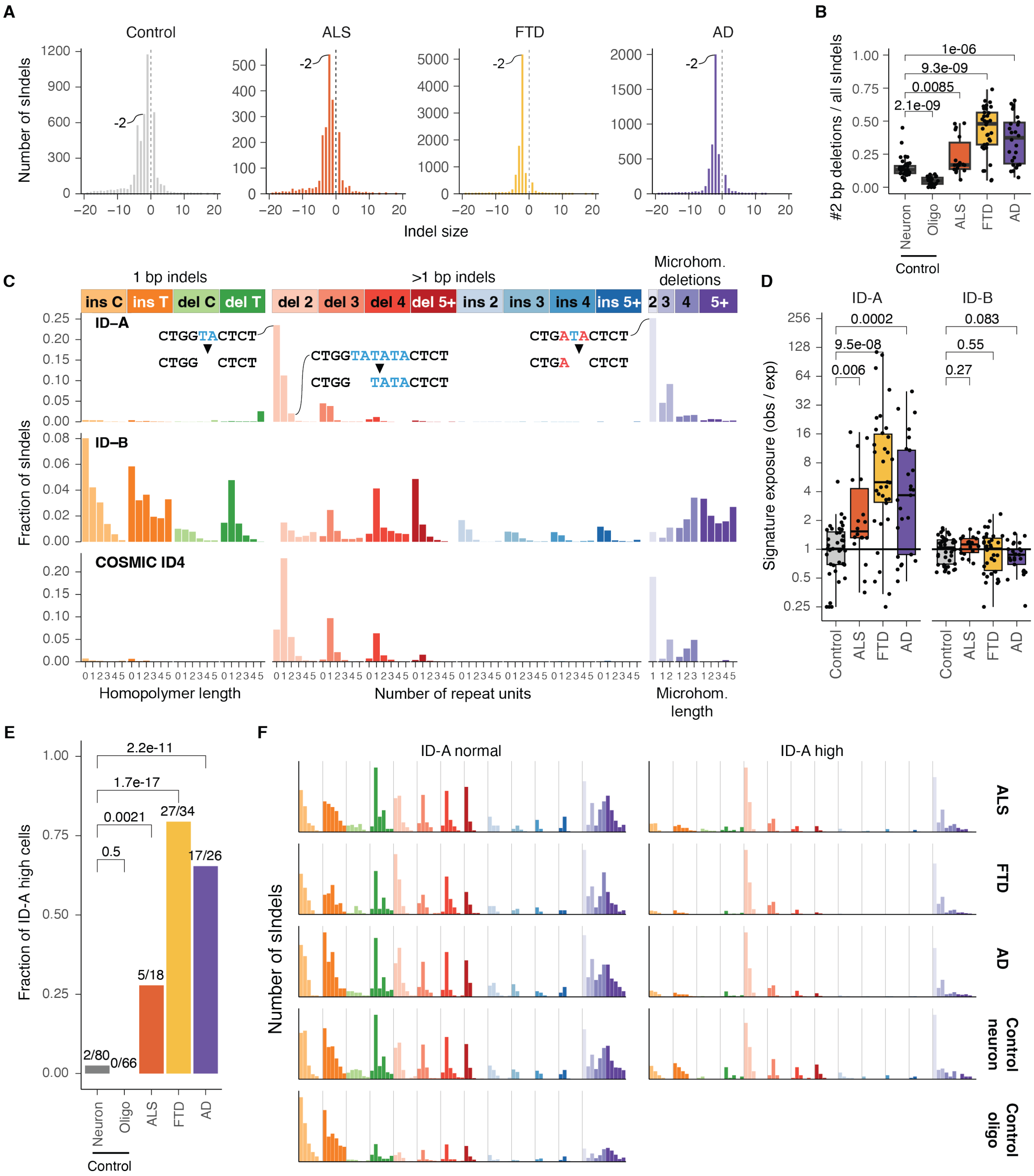

Fig. 3. Neurons from *C9ORF72* ALS, *C9ORF72* FTD and AD brains share an indel mutational signature. (**A**) Distribution of sIndel lengths across conditions. Diseased neurons have higher levels of 2-bp deletions. (**B**) Fraction of 2 bp deletions in neurons and oligodendrocytes across conditions. All P-values in this figure represent Wilcoxon rank-sum tests. (**C**) Both *de novo* sIndel signatures identified from joint analysis of diseased and

neurotypical control neurons amplified by PTA, as well as previously published neurons and oligodendrocytes amplified using an older amplification technology (MDA). COSMIC ID4 is a signature previously identified in cancer and normal aging neurons. Top part illustrates the disease-associated 2-bp deletions in ID-A signature. (**D**) Burdens of ID-A and ID-B sIndels; P-values are Wilcoxon rank-sum tests. (**E**) Fraction of cells with high levels of ID-A sIndels in neurons and oligodendrocytes across conditions, determined by a classifier (Methods); P-values represent Fisher tests of independence between disease status and ID-A-high classifications. (**F**) sIndel spectra of neurons and oligodendrocytes stratified by ID-A-high and -normal classifications. Oligo: oligodendrocyte.

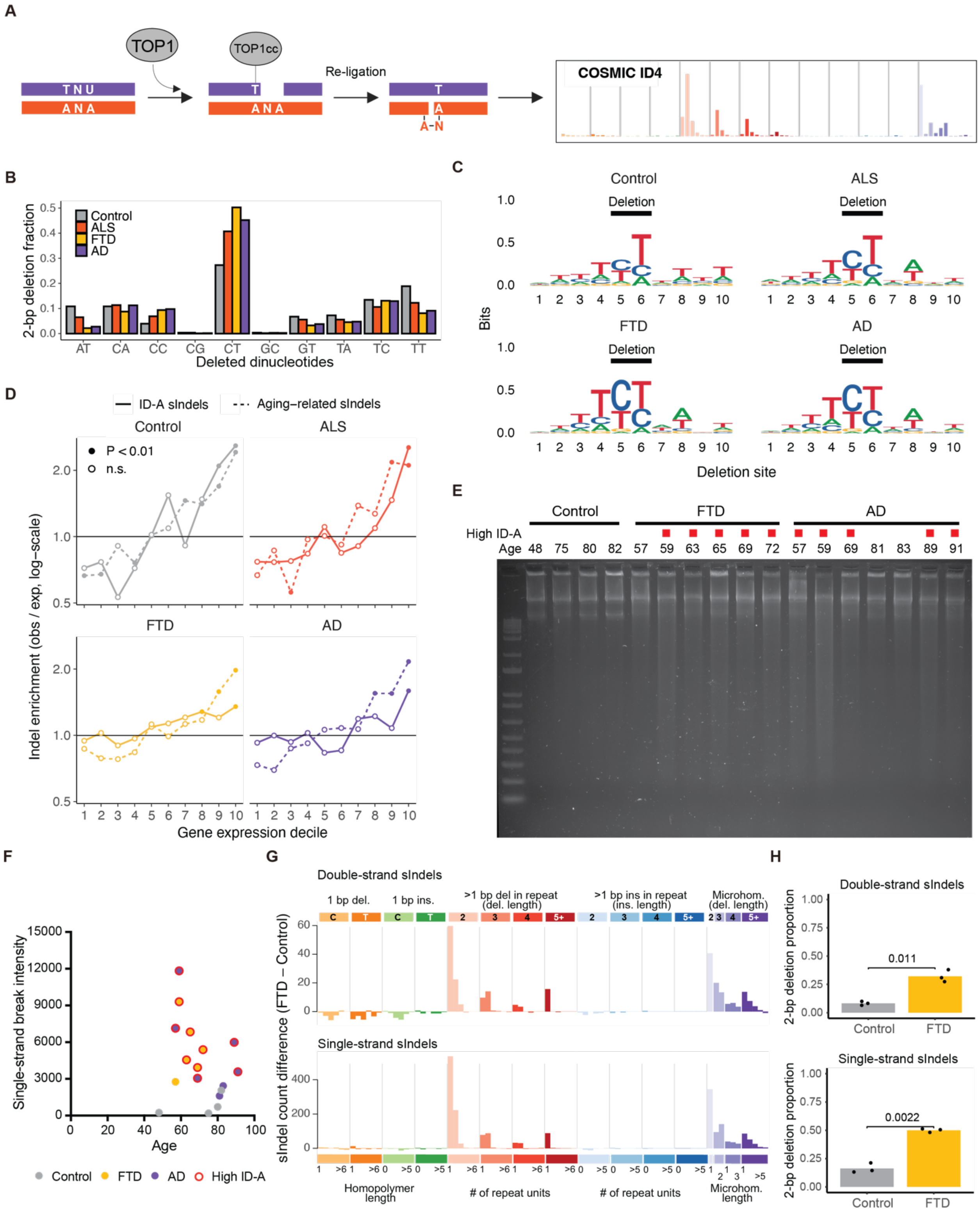

Fig. 4. ID-A-like deletions reflect characteristics of TOP1-mediated ribonucleotide excision. (**A**) Model of TOP1-mediated ribonucleotide excision repair. (**B**) Rates of 2-bp deletions by the

deleted dinucleotide. (**C**) DNA motifs of 2-bp deletion sites; positions 5 and 6 are the deleted nucleotides. (**D**) Enrichment of sIndels in relation to local gene expression levels from previously published snRNA-seq data of excitatory neurons. Enrichment tests divide the genome into 10 roughly equally-sized, non-contiguous regions ranked by gene expression. I.e., decile 10 represents the subset of the genome containing the top 10% of expressed genes. Solid lines: enrichment analysis of ID-A-like sIndels (see Fig. 3A), dotted lines: all other sIndels. P-values are enrichment tests (Methods) indicating that the observed/expected ratio significantly differs from 1. (**E**) Agarose gel electrophoresis of denatured gDNA from prefrontal cortex tissues, ordered by age. Smears on the gel indicate single-strand breaks. Red rectangles indicate cases with high ID-4 like sIndels (Methods). (**F**) Abundance of single-strand breaks in prefrontal cortex tissues. Y-axis values are quantifications of the smears in (**E**). Red circles indicate brains for which most scWGS neurons were classified as ID-A-high. (**G**) Indel spectra of excess (subtracting sIndels of age-matched controls) double-strand and single-strand sIndels in neurons from *C9ORF72* FTD cases compared to neurons from neurotypical controls. Mini-bulk samples (50 neurons per case) were sequenced using META-CS. (**H**) Fraction of 2-bp deletions among single-strand lesions and double-strand sIndels (numbers of 2-bp deletions divided by total numbers of Indels). P-values are two-tailed unpaired *t*-test.

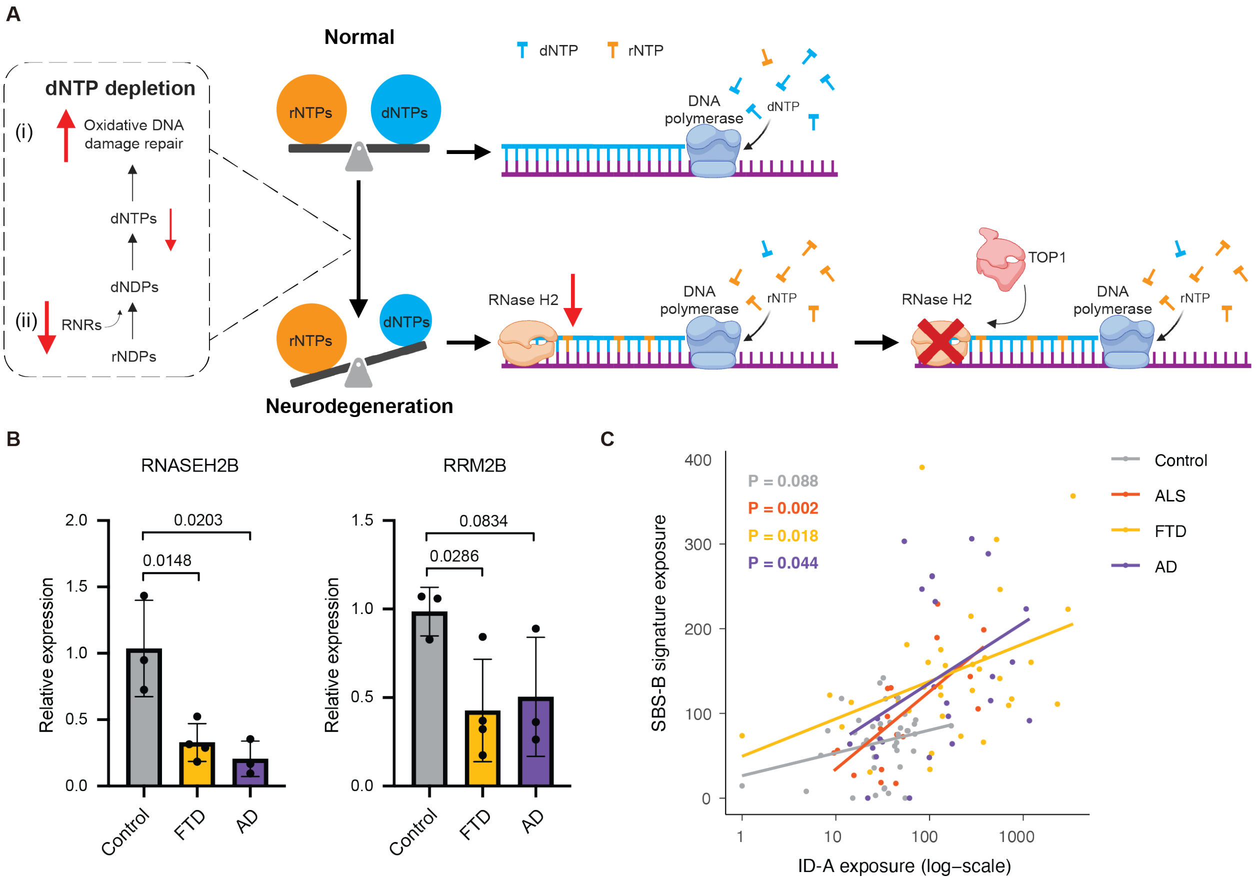

Fig. 5. Potential mechanisms underlying the dysregulation of ribonucleotide excision repair. (**A**) Model of TOP1-mediated mutagenesis in neurons from neurodegenerative conditions. RNR: ribonucleotide reductase. (**B**) Expression levels of *RNASEH2B* and *RRM2B* in neurons measured by qPCR. P-values are two-tailed unpaired *t*-test. (**C**) Correlation between burdens of *de novo* signatures SBS-B sSNVs and ID-A sIndels. SBS-B resembles the COSMIC SBS30 signature, which is associated with a deficiency in NTHL1 and base excision repair of oxidative DNA damage repair. P-values indicate regression t-tests on slope.

Methods

Tissue sources

Fresh frozen postmortem human brain tissues were originally collected by the Massachusetts Alzheimer's Disease Research Center and NIH NeuroBioBank (table S1) according to their respective institutional protocols, written authorization and informed consent; they were subsequently obtained for this study with the approval of the Boston Children's Hospital Institutional Review Board. Research on these deidentified specimens and data was performed at Boston Children's Hospital with approval from the Committee on Clinical Investigation. The presence of *C9ORF72* repeat expansion in these cases were determined through repeat-primed PCR (RP-PCR).

Nuclei isolation and single-cell whole genome amplification

Single neuronal nuclei were isolated using FANS with co-staining of NeuN and TDP-43 antibodies, as described previously(*26, 60*). Briefly, fresh frozen postmortem human brain tissues were homogenized in a dounce homogenizer with a chilled tissue lysis buffer (10mM Tris-HCl, 0.32M sucrose, 3mM Mg(Oac)2, 5mM CaCl2, 0.1mM EDTA, 1mM DTT, 0.1% Triton X-100, pH 8) on ice. Tissue lysates were carefully overlayed on top of a sucrose cushion buffer (1.8M sucrose 3mM Mg(Oac)2, 10mM Tris-HCl, 1mM DTT, pH 8) and ultra-centrifuged for 1 hour at 30,000 x g. Nuclear pellets were incubated and resuspended in ice-cold PBS supplemented with 3mM MgCl2, filtered (40 µm pore size), then stained with Alexa Fluor 488-conjugated anti-NeuN antibody (Millipore MAB377X) and Alexa Fluor 647-conjugated anti-TDP-43 antibody (Abnova H00023435-M01). DAPI was added to stained nuclei right before the sorting. Single neuronal nuclei with TDP-43 pathology (NeuN+, TDP-43-) and without TDP-43 pathology (NeuN+, TDP-43+) were then sorted into individual wells of 96-well plates. Primary template-directed amplification (PTA) of sorted single nuclei was then performed using the ResolveDNA® Whole Genome Amplification Kit v1 (BioSkryb 100136) following the manufacturer's protocol as previously described(*25*). Amplified single-cell genomes were subjected to quality control with Picogreen and a 4-locus multiplex PCR(*60*). Single-cell genomes with sufficient DNA yield and successful amplification of all four loci were selected for scWGS.

Library preparation for scWGS

Libraries were made from PTA amplified single-cell genomes following a modified KAPA HyperPlus Library Preparation protocol provided by BioSkryb as previously described(*25*). Libraries were subjected to quality control using Picogreen and Tapestation HS D1000 Screen Tape (Agilent PN 5067–5584) before sequencing. Single cell genome libraries were sequenced on the Illumina NovaSeq platform (150 bp x 2) at 30X.

scWGS data preprocessing

Illumina reads were aligned to the human reference with decoy sequence GRCh37d5 (hs37d5) using bwa mem version 0.7.15-r1140 with the -M option. Following alignment, BAMs were processed according to the GATK Best Practices recommendations, namely Picard MarkDuplicates version 2.8.0 and IndelRealigner, BaseRecalibrator and PrintReads using GATK version 3.5.0-g36282e4.

scWGS somatic mutation calling with SCAN2

Somatic SNVs and indels were called from PTA and MDA data using SCAN2 version 1.2.13, git commit `09ed5b9` and r-scan2 git commit `c0aef34`. The standard SCAN2 recommended workflow was followed. First, `scan2 makepanel` was run on the full catalog (i.e., across all disease statuses, amplification techniques, cell types, etc.) of PTA and MDA single cells and matched bulks. The cross-sample panel produced by this step is required for somatic indel calling. Second, one instance of `scan2 call_mutations` was run for each unique individual. For each run, all BAMs belonging to the individual (i.e., all MDA, PTA and matched bulk(s)) were provided to SCAN2. Third, one instance of SCAN2's mutation-signature based rescue (`scan2 rescue`) was run for each unique combination of amplification type (PTA or MDA), phenotype (ALS, FTD, AD or control) and cell type (neuron, oligodendrocyte or mixed glia). Finally, SCAN2's post-processing script to filter clustered or recurrent (i.e., called in >1 single cell) artifacts and mutations (`SCAN2/bin/digest_calls.R`) was applied to each `scan2 rescue` output and only calls with `final.filter=FALSE` were retained. Critically, because of the unique mutational signature of disease-associated indels, we opted to exclude all SCAN2 rescue mutation calls from all analyses to avoid any bias involved in signature-based rescue. Rescued calls were removed by excluding rows with `rescue=TRUE` from the final recurrence filtered output. Additional configuration options applicable to all SCAN2 steps were `--gatk=sentieon_joint`; the GRCh37 human reference genome with decoy hs37d5 (`--ref`), dbSNP v147 common (`--dbsnp`) and 1000 Genomes phase 3 SHAPEIT2 phasing panel (`--shapeit-refpanel`).

scWGS mutation burden estimation and comparison

SCAN2 estimates the genome-wide mutation burden by correcting the number of called mutations for detection sensitivity. This correction accounts for sensitivity changes caused by sequencing depth, imbalanced and/or non-uniform single-cell amplification, and other factors. Correction is performed separately for sSNVs and sIndels. More details are provided in our previous publication(*25*). SCAN2's sSNV and sIndel mutation burden estimates were extracted from each SCAN2 object in R using r-scan2's `mutburden()` function. To compare mutational burdens between any two groups (e.g., control and ALS), it is first necessary to correct for differences in age between the two groups. Two strategies were used together to achieve this goal. First, we computed the expected burden as a function of age using PTA-amplified, neurotypical control neurons. This process involved two steps of linear models; the first to identify strong outliers and the second to compute the final age-based expectation. In more detail, the first linear model was fit to all PTA control neurons via `lm(burden ~ age)` and Tukey's method was applied to the model's residuals to identify strong outliers. The second linear model was then fit with the outliers removed and R's `predict` function was used to derive the expected burden as a function of age. Separate models were fit to sSNV, sIndel and mutational signature burdens (see *Mutational signature analysis* for more details on signature burdens); for analysis of oligodendrocytes, n=66 PTA-amplified control oligodendrocytes were used in place of PTA control neurons. For each cell and burden type (i.e., sSNV, sIndel, etc.) a final observed/expected ratio was produced by dividing the observed burden from SCAN2 by the expected burden based only on the cell's age from the linear model. Second, for all burden comparisons between groups, only observed/expected ratios from cells with age > 50 were included since the residuals of younger cells tended to be large compared to the trend line, leading to inflated observed/expected ratios. Significance tests for differences in burdens between groups are Wilcoxon rank-sum tests on the observed/expected ratios.

scWGS quality control and removal of outliers

To ensure that technical factors did not drive differences in burdens or mutational signatures, we compared sSNV and sIndel burdens against a quantity that reflects quality of single cell amplification, the median absolute pairwise difference (MAPD). Low MAPD values indicate more uniform amplification and thus better quality. Plots of mutation burden vs. MAPD were visually assessed to identify cells in which technical quality may have driven high sIndel burdens (i.e., cells with both high sIndel and high MAPD). This revealed 5 PTA-amplified neurons (3 AD, 2 FTD) and 7 MDA-amplified neurons (6 AD, 1 control), which were excluded from further analysis. Additionally, 10 MDA-amplified neurons were excluded due to extreme sIndel burden and/or a highly atypical mutation signature enriched in long microhomology deletions.

Mutational signature analysis

Matrices of SBS96 and ID83 somatic mutation counts were generated with SigProfilerMatrixGenerator version 1.2.26 using the GRCh37 human reference genome. All single cells (both MDA- and PTA-amplified; table S2), including outliers, were included for matrix generation. Next, sIndel count matrices were corrected to reflect different sensitivities for the 83 indel channels, as estimated in our previous publication(*25*). De novo mutational signatures were then extracted using SigProfilerExtractor version 1.1.24 and options min_stability=0.9, minimum_signatures=1, maximum_signatures=8, and nmf_replicates=100. The decompositions with the most stable signatures were selected, which was K=6 signatures for SBS96 and K=2 signatures for ID83. For each de novo signature, a genome-wide burden was computed by multiplying the SigProfilerExtractor assigned number of mutations by SCAN2's genome-wide scaling factor for the appropriate mutation type (i.e., SBS96 counts were multiplied by SCAN2's sSNV genome-wide burden scaling factor and ID83 counts were multiplied by SCAN2's sIndel genome-wide scaling factor). These scaled genome-wide burdens were then used to fit models on PTA-amplified control neurons and calculate age-adjusted expected values using the procedure detailed in *scWGS mutation burden estimation and comparison.*

Classification of ID-A high and normal cells

Each cell was classified as either ID-A normal (expected levels of ID-A) or ID-A high (excessive ID-A, consistent with disease progression). To account for data sparsity in individual cells, indel spectra were summarized by two numbers that represent the primary features of ID-A: the fraction of 2 bp deletions without microhomology (which corresponds to the pink-colored peaks in the ID83 spectrum) and the fraction of 2-3 bp deletions with microhomology (which corresponds to the light purple peaks in the ID83 spectrum). Further, cells with fewer than 20 indels were excluded from classification. Clusters were identified in this 2-dimensional space using the `mclustBIC` function from the `mclust` R package, which produced 4 clusters. Clusters were visually assessed and two clusters identified by high rates of both forms of deletion characteristic to ID-A were labeled ID-A high. All other clusters were labeled ID-A normal.

Mutation enrichment analysis

Enrichment tests. Enrichment analysis was performed following the procedures outlined in our previous publications(*25, 34*). Briefly, enrichment analysis considers the number of somatic mutations over a set of genomic regions which are not, in general, contiguous (e.g., defined by covariates such as gene expression). The null hypothesis is that somatic mutations are uniformly distributed over the genome. To test the null hypothesis, SCAN2 simulates uniformly distributed mutations across the genome for each single cell, considering both (1) the regions of the genome with sufficient read depth to detect somatic mutations and (2) the mutational spectrum of the whole mutation set, which is required to match the observed somatic mutations. This simulation is performed 10,000 times to produce 10,000 null sets of mutations. The true mutation sets and null sets of mutations are then combined for each set of single cells of interest (e.g., all PTA single neurons from ALS). For each genomic region, the true mutation set and simulation sets are intersected with the region and tallied. The enrichment over a region is defined as the number of true mutations in the region divided by the mean number of simulated mutations in the region over the 10,000 simulations; values > 1 indicate enrichment while values < 1 indicate depletion. This same metric is be applied to all 10,000 simulations to construct a distribution of the enrichment statistic. A two-sided enrichment test *P*-value is constructed by counting the number of simulations with more extreme enrichment ratios than the observed mutation count (to make this test two-sided, more extreme is defined as having greater absolute log(enrichment)).

Enrichment regions. Seven diverse region definitions were used to assess mutation enrichment. (1) The ChromHMM 15-state model from dorsolateral prefrontal cortex (ID: E073), which assigns every genomic location one of 15 states(*43*); (2) transcribed and untranscribed genomic regions as determined by the GENCODEv26 gene model (UTRs, exons and introns were defined as transcribed); (3) promoters and enhancers active in neurons(*45*); (4) ChIP-seq measurements of histone modifications H3K27ac, H3K4me1, H3K4me3, H3K9ac, H3K9me3 from the Roadmap Epigenomics Project(*43*); (5) RNA-sequencing gene expression levels in the prefrontal cortex (BA9) from the GTEx Portal, v8 release(*61*); (6) single nucleus RNA-sequencing gene expression levels from excitatory neurons;(*34*) and (7) single nucleus ATAC-sequencing of accessible chromatin in excitatory neurons(*34*).

Single-nucleus RNA sequencing library preparation

For each of the premotor cortex samples of *C9ORF72* ALS brains and prefrontal cortex samples of *C9ORF72* FTD brains, ten thousand neuronal nuclei with TDP-43 pathology (NeuN+, TDP-43-) and without TDP-43 pathology (NeuN+, TDP-43+) were sorted into individual wells of 96-well plates. For each of the premotor cortex samples and prefrontal cortex samples of age-matched normal brains, only ten thousand neuronal nuclei without TDP-43 pathology (NeuN+, TDP-43+) were sorted. Sorted nuclei were then used for droplet generation and sequencing library preparation using the 10X Genomics Next GEM Single Cell 3′ GEM Kit v3.1 and Chromium Controller, following the manufacturer's manual.

Cryptic exon junction analysis

Annotated TDP-43 dependent cryptic junctions were obtained from the previous study and converted from hg38 to hg19 genome assembly(*27*). Cryptic exon junction analysis was performed following the recent study with some modifications(*28*). Briefly, BAM files generated by cellranger were filtered to exclusively include reads meeting the criteria set in the snRNA-seq analysis (described above). For neuron type-specific junction analysis, BAM files were split into distinct neuron type-specific BAM files using each cell’s 10X barcode and respective cluster annotation. These filtered BAM files were then subjected to analysis using regtools junctions extract, employing parameters “-a 6 -m 30 -M 500000 -s RF” to product junction files(*62*). The junctions were then subjected to Leafcutter’s leafcutter_cluster_regtools.py by employing “-m 10

-p 0.0001" to generate a summary matrix of counts of all junctions found in the dataset(*63*).Given the limited number of supporting reads for splicing junctions, the sum of numbers of reads from two replicates was used. Only genes with at least 3 reads supporting the splicing events were analyzed (sum of the number of constitutive splicing reads and the number of cryptic exonization reads; the denominator of regtools output).

META-CS library preparation for mini-bulk samples of neuronal nuclei

META-CS libraries were made following the previously described protocol with modifications (DOI: http://dx.doi.org/10.17504/protocols.io.6qpvr3nbzvmk/v1)(*46*). Briefly, 50 neuronal nuclei were sorted into individual wells of 96-well plates containing 2 µL of META lysis buffer (20 mM Tris, pH 8.0, 20 mM NaCl, 0.15% Triton X-100, 25 mM dithiothreitol, 1 mM EDTA, 3 units/mL NEB Thermolabile Proteinase K [P8111S]) at 30 °C for 1 h, 55 °C for 10 min. Lysed cells were stored at -80 °C until library preparation.

The transposome was assembled using Tn5 transposase (Diagenode C01070010) following manufacturer's manual (5 µL of annealed META mix, 5 µL of Tn5, 5 µL of glycerol). The assembled transposome was further diluted using the Tagmentase Dilution Buffer (Diagenode C01070011) based on optimized Tn5 concentration (performed on the day of library preparation). After thawing on ice, cell lysate was transposed with 8 µL of transposition mix (1 µL of diluted transposome, 5 µL of 2X Tagmentation Buffer [Diagenode C01019043], 2 µL of water) and incubated at 55 °C for 15 min. The transposition was stopped by adding 2 µL of 6X stop buffer (300 nM NaCl, 45 mM EDTA, 0.01% Triton X-100, , 3 units/mL Thermolabile Proteinase K [NEB P8111S]) and incubated at 37 °C for 30 min, 55 °C for 10 min.

First-strand tagging was performed by adding 13 µL of Strand Tagging Mix 1 (5 µL of Q5 reaction buffer, 5 µL of Q5 high GC enhancer, 0.85 µL of 100 µM [total] Adp1 primer mix, 0.6 µL of 100 mM MgCl2, 0.6 µL of water, 0.5 µL of 10 mM dNTP mix, 0.25 µL of 20 mg/mL bovine serum albumin [NEB B9200], 0.25 µL of Q5 DNA polymerase [NEB M0491]) and incubated at 72 °C for 3 min, 98 °C for 30 s, 62 °C for 5 min, 72 °C for 1 min. Adp1 primers were removed by adding 1 µL of Thermolabile ExoI (NEB M058) and incubated at 37 °C for 15 min, 65 °C for 5 min.

Second-strand tagging was performed by adding of 4 µL of Strand Tagging Mix 2 (1 µL of Q5 reaction buffer, 1 µL of Q5 high GC enhancer, 0.95 µL of 100 µM [total] Adp2 primer mix ,

0.945 µL of water, 0.1 µL of 10 mM dNTP mix, 0.05 µL of Q5 DNA polymerase) and incubated at 98 °C for 30 s, 62 °C for 5 min, 72 °C for 1 min. Adp2 primers were removed by adding 1 µL of Thermolabile ExoI (NEB M058) and incubated at 37 °C for 15 min, 65 °C for 5 min.
Strand tagging products were amplified by adding 19 µL of PCR mix (5 µL of NEBNext Multiplex Oligos Universal Primer, 5 µL of NEB Index Primers [NEB E7335S, E7500S, E7710S, E7730S], 4 µL of Q5 reaction buffer, 4 µL of Q5 high GC enhancer, 0.4 µL of 10 mM dNTP mix, 0.4 µL of water, 0.2 µL of Q5 DNA polymerase) and incubated at 98 °C for 20 s, 11 cycles of (98 °C for 10 s, 72 °C for 2 min), 72 °C for 2 min.
Size selection of META-CS libraries were performed using AMPure XP beads (Beckman Coulter A63882) with a 0.65X upper cut and a 1.8X lower cut following manufacturer's manual. Libraries were eluted into 30 µL of low TE buffer. Each META-CS library was sequenced on one lane of the Illumina HiSeq X sequencer.

Denaturation of gDNA and agarose gel electrophoresis

gDNA were extracted from the prefrontal cortex of *C9ORF72* FTD, AD and control brains using the Qiagen EZ1 Advanced XL system (Qiagen 9001874). Each gDNA sample (400 g) was diluted to a final volume of 50 µL with water. gDNA samples were then concentrated using ethanol precipitation with glycogen as a carrier. To each gDNA sample, 0.1 volumes of 3 M sodium acetate, 20 ng of glycogen, and 3 volumes of ethanol were added. The mixture was gently resuspended and incubated at –80°C for 1 hr. Samples were centrifuged at 12,000 × g for 30 minutes at 4°C, and the supernatant was carefully removed. The DNA pellet was washed with chilled 70% ethanol, air-dried, and dissolved in 2 µL of nuclease-free water. For denaturation, 27 µL of formamide and 1 µL of 0.5 M EDTA (pH 8.0) were added to the dissolved DNA, and the mixture was incubated at 37°C for 1 hour to generate single-strand DNA. After adding 6 µL of 6X gel loading buffer, the denatured DNA samples were separated on 1% agarose gels. Gels were stained with SYBR Gold and imaged using a gel scanner. The intensity of the smears (single-strand breaks) was measured using ImageJ. Cases with high or normal levels of ID-A sIndels were classified by the fractions of neurons with high levels of ID-A sIndels.

Preprocessing of META-CS data

We preprocessed the META-CS data using a pipeline that expands the previously reported workflow (Xing 2021 PNAS) with additional features to accommodate both our modified experimental protocol and the new somatic indel calling method. First, paired-end reads were preprocessed by customized pre-meta which identifies Tn5 barcodes, merges overlapping read ends, and trims Illumina adapters. Then, BWA-MEM (v.0.7.17)(*64*) was used to map reads to the human reference genome (GRCh37 with decoy) and generate a BAM file for each sample. Importantly, we then split the BAM file by unique pairs of barcodes which are used to distinguish original DNA fragments and help ensure that mutation calling was performed on a single-molecule level. Furthermore, given a relatively small pool of unique barcodes, barcode collision may occur where different DNA fragments are tagged by the same barcode pair around the same genomic region. This possibility also increases with our pooled-cell input with many more available DNA molecules at any locus. To address this challenge, we extracted Tn5 cut sites of each molecule (i.e. the start and end positions of each read pair) and used a combination of shared Tn5 barcode pair and Tn5 cut sites to identify original DNA fragments.

Somatic indel calling from META-CS data

As the original META-CS method was developed for somatic SNVs, we established a new somatic indel calling pipeline that introduces novel modules to enhance the calling accuracy and remove false positives specific for indels (manuscript in preparation). First, we followed a similar workflow as SNV calling and piled up reads from the META-CS barcode pair BAMs and a matching bulk WGS BAM to generate somatic indel candidates by identifying variants that have at least 4 total non-reference (ALT) reads in META-CS with at least 2 ALT reads from each strand for duplex support, as well as no ALT read in bulk to remove potential germline variants. Since the majority of false positive indels derive from incorrect read merging during preprocessing, we created modules to tag the merged reads with the window where merging occurred and generate an additional BAM file without read merging. Then, we filtered out candidates that are either within the merging window or not present in the BAM without read merging. In addition, we implemented a set of filters. A candidate site was filtered out if it met any of the following criteria, 1) overlapping with the low-quality regions as previously reported (um75- hs37d5.bed(*46*)), 2) overlapping with gnomAD indels of ≥ 1% population frequency, 3) located within 100 bp from another candidate site, 4) located within 10 bp of read ends, 5) having multiple Tn5 cut sites.

Single-strand indel calling from META-CS data

Similar to double-strand indel calling described above, single-strand indel candidates were generated by piling up reads from the META-CS barcode pair BAMs and a matching bulk WGS BAM, followed by identifying variants that have no ALT read from bulk and at least 8 total reads in META-CS with at least 4 reference allele (REF) reads from one strand (i.e. non-variant strand) and at least 4 ALT reads from the other strand (i.e. variant strand). In addition, the candidate is required to have no ALT read from the non-variant strand and no REF read from the variant strand. Other filters are the same as double-strand calling above.

ALS and FTD snRNA-seq analysis

For snRNA-seq analysis on ALS and FTD, we used data generated for this study (i.e. in-house). We implemented the following pipeline to raw FASTQ files from both data sets. Cellranger (v.6.1.0)(*65*) with “--include-introns” was used to map reads to the human reference genome (GRCh37) and generate gene-by-cell count matrices. To remove potential ambient RNA contamination, we ran remove-background from CellBender (v.0.3.0)(*66*)on the raw count matrices with the default parameters to generate filtered count matrices for downstream analysis. For quality control, we filtered out cells with < 700 genes and mitochondrial gene percentage > 20%, which resulted in more than 65% cells removed in one sample (1700BA6-normal). So this sample was excluded for further analysis. scDblFinder(*67*)was used to remove doublets, and DropletQC(*68*)was used to calculate nuclear fractions where cells with a nuclear fraction < 0.25 were filtered out. Next, we used Seurat (v.5.1.0)(*69*) SCTransform v2 to perform normalization and variance stabilization followed by RunHarmony to integrate and harmonize cell embeddings across all samples. The harmonized embeddings were used to perform dimension reduction (RunUMAP) and clustering using the Leiden algorithm(*70*). Cell types were annotated using previously reported markers(*71*)

Cell type fraction analysis in snRNA-seq

We used a Bayesian statistics framework(*72, 73*) to detect changes in cell type fractions in ALS and FTD compared to controls. This model implements Dirichlet multinomial modeling – Hamiltonian Monte Carlo (DMM-HMC) using R packages rstan (v.2.32.6) and bayestestR (v0.13.2) (http://mc-stan.org/). The input was a count matrix where each row is a sample and each column is a cell type. Counts in each row of the matrix follow a multinomial distribution where proportions of each cell type in the sample are modeled by Dirichlet. The prior probability of the proportions is another Dirichlet distribution with the fixed parameter alpha = 10^{-7} for equal expected fractions of cell types. We split samples into groups based on their disease status, brain region, and TDP43 status, and calculated the posterior probability distribution (PPD) of each group. Then, we subtracted the PPD of control group from that of disease group to detect any significant changes in cell type fractions. Significance was determined if the log2 ratio of disease/control has 95% credible intervals above or below zero.

Differential gene expression analysis in snRNA-seq data

Differential gene expression analysis was performed using Nebula (v.1.5.3)(*74*) in R on in-house ALS and FTD data set and AD data set from SEA-AD consortium(*75*). We downloaded the processed AD snRNA-seq data from SEA-AD consortium's web portal (sea-ad.org). For each data set, raw count matrices after preprocessing and cell type annotation were used as inputs and the design matrix was constructed with disease status as the predicator. Then, a negative binomial mixed model was built for each gene using individual as the random effect and total number of molecules in each cell as offsets. The output raw p-values were adjusted using the Benjamini Hochberg method. Differentially expressed genes with FDR < 0.05 were considered significant.

qPCR analysis of RNASEH2 and RRM gene expression

Neuronal nuclei (2350 NeuN+ nuclei, ~5 µL) from *C9ORF72* FTD, AD and age-matched control brains were sorted into individual wells of 96-well plates with triplicates. Cell lysis, reverse transcription and qPCR were conducted using the Power SYBR™ Green Cells-to-CT™ Kit (Thermo Fisher 4402954) following the manufacturer's protocol. The following primers were used for qPCR analysis:

hRNASEH2A Forward - 5'- ACAGCCACTGGGCTTATACAG -3' hRNASEH2A Reverse - 5'- TCCCTACGGTGTCCACGAATA -3'

hRNASEH2B Forward - 5'- TAACCCCTGTTCAGGAGAAGG -3'

hRNASEH2B Reverse - 5'- ACACGTTATCCACCACAACTTG -3'

hRNASEH2C Forward - 5'- AGGGACTCGAAGTGTCGTTTC -3'

hRNASEH2C Reverse - 5'- TGTCACCATCACGTATCCCAC -3'

hRRM1 Forward - 5'- GCCAGGATCGCTGTCTCTAAC -3'

hRRM1 Reverse - 5'- GAGAGTGTTTGCCATTATGTGGA -3'

hRRM2 Forward - 5'- ATTGGGCCTTGCGATGGATAG -3'

hRRM2 Reverse - 5'- GAGTCCTGGCATAAGACCTCT -3'

hRBFOX3 Forward - 5'- TCGTAGAGGGACGGAAAATTGA -3'

hRBFOX3 Reverse - 5'- GCCGTTGGTGTAGGGGTTC -3'

References

1. T. Niccoli, L. Partridge, A. M. Isaacs, Ageing as a risk factor for ALS/FTD. *Hum Mol Genet* **26**, R105-R113 (2017).

2. L. E. Hebert *et al.*, Age-specific incidence of Alzheimer's disease in a community population. *JAMA* **273**, 1354-1359 (1995).

3. S. C. Ling, M. Polymenidou, D. W. Cleveland, Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* **79**, 416-438 (2013).

4. M. DeJesus-Hernandez *et al.*, Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245-256 (2011).

5. A. E. Renton *et al.*, A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257-268 (2011).

6. B. T. Hyman *et al.*, National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement* **8**, 1-13 (2012).

7. R. J. Ferrante *et al.*, Evidence of increased oxidative damage in both sporadic and familial amyotrophic lateral sclerosis. *J Neurochem* **69**, 2064-2074 (1997).

8. A. R. Haeusler *et al.*, C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* **507**, 195-200 (2014).

9. J. Mitra *et al.*, Motor neuron disease-associated loss of nuclear TDP-43 is linked to DNA double-strand break repair defects. *Proc Natl Acad Sci U S A* **116**, 4696-4705 (2019).

10. C. D. Smith *et al.*, Excess brain protein oxidation and enzyme dysfunction in normal aging and in Alzheimer disease. *Proc Natl Acad Sci U S A* **88**, 10540-10543 (1991).

11. E. Adamec, J. P. Vonsattel, R. A. Nixon, DNA strand breaks in Alzheimer's disease. *Brain Res* **849**, 67-77 (1999).

12. S. P. Gabbita, M. A. Lovell, W. R. Markesbery, Increased nuclear DNA oxidation in the brain in Alzheimer's disease. *J Neurochem* **71**, 2034-2040 (1998).

13. M. A. Lodato *et al.*, Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555-559 (2018).

14. M. B. Miller *et al.*, Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714-722 (2022).

15. D. R. Rosen *et al.*, Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362**, 59-62 (1993).

16. K. Acs *et al.*, The AAA-ATPase VCP/p97 promotes 53BP1 recruitment by removing L3MBTL1 from DNA double-strand breaks. *Nat Struct Mol Biol* **18**, 1345-1350 (2011).

17. M. Meerang *et al.*, The ubiquitin-selective segregase VCP/p97 orchestrates the response to DNA double-strand breaks. *Nat Cell Biol* **13**, 1376-1382 (2011).

18. W. Y. Wang *et al.*, Interaction of FUS and HDAC1 regulates DNA damage response and repair in neurons. *Nat Neurosci* **16**, 1383-1391 (2013).

19. R. Lopez-Gonzalez *et al.*, Poly(GR) in C9ORF72-Related ALS/FTD Compromises Mitochondrial Function and Increases Oxidative Stress and DNA Damage in iPSC-Derived Motor Neurons. *Neuron* **92**, 383-391 (2016).

20. G. Hewitt *et al.*, SQSTM1/p62 mediates crosstalk between autophagy and the UPS in DNA repair. *Autophagy* **12**, 1917-1930 (2016).

21. J. Higelin *et al.*, NEK1 loss-of-function mutation induces DNA damage accumulation in ALS patient-derived motoneurons. *Stem Cell Res* **30**, 150-162 (2018).

22. S. Cohen *et al.*, Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat Commun* **9**, 533 (2018).

23. M. Maor-Nof *et al.*, p53 is a central regulator driving neurodegeneration caused by C9orf72 poly(PR). *Cell* **184**, 689-708 e620 (2021).

24. M. Mirceta *et al.*, C9orf72 repeat expansion creates the unstable folate-sensitive fragile site FRA9A. *NAR Mol Med* **1**, ugae019 (2024).

25. L. J. Luquette *et al.*, Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat Genet* **54**, 1564-1571 (2022).

26. E. Y. Liu *et al.*, Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. *Cell Rep* **27**, 1409-1421 e1406 (2019).

27. X. R. Ma *et al.*, TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. *Nature* **603**, 124-130 (2022).

28. L. M. Gittings *et al.*, Cryptic exon detection and transcriptomic changes revealed in single-nuclei RNA sequencing of C9ORF72 patients spanning the ALS-FTD spectrum. *Acta Neuropathol* **146**, 433-450 (2023).

29. J. P. Ling, O. Pletnikova, J. C. Troncoso, P. C. Wong, TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science* **349**, 650-655 (2015).

30. M. Nolan *et al.*, Quantitative patterns of motor cortex proteinopathy across ALS genotypes. *Acta Neuropathol Commun* **8**, 98 (2020).

31. B. R. Foerster *et al.*, An imbalance between excitatory and inhibitory neurotransmitters in amyotrophic lateral sclerosis revealed by use of 3-T proton magnetic resonance spectroscopy. *JAMA Neurol* **70**, 1009-1016 (2013).

32. C. S. Khademullah *et al.*, Cortical interneuron-mediated inhibition delays the onset of amyotrophic lateral sclerosis. *Brain* **143**, 800-810 (2020).

33. I. Allodi, R. Montanana-Rosell, R. Selvan, P. Low, O. Kiehn, Locomotor deficits in a mouse model of ALS are paralleled by loss of V1-interneuron connections onto fast motor neurons. *Nat Commun* **12**, 3251 (2021).

34. J. Ganz *et al.*, Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. *Cell* **187**, 1955-1970 e1923 (2024).

35. J. M. Charcot, De la sclerose laterale amyotrophique. Symptomatologie. *Lecons sur les Maladies du Systeme Nerveux Faites a la Salpetriere* **2**, 227-242 (1880).

36. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, M. R. Stratton, Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-259 (2013).

37. M. Petljak *et al.*, Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294 e1220 (2019).

38. A. Klungland *et al.*, Base excision repair of oxidative DNA damage activated by XPG protein. *Mol Cell* **3**, 33-42 (1999).

39. M. A. M. Reijns *et al.*, Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature* **602**, 623-631 (2022).

40. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).

41. A. Tanizawa, K. W. Kohn, Y. Pommier, Induction of cleavage in topoisomerase I c-DNA by topoisomerase I enzymes from calf thymus and wheat germ in the presence and absence of camptothecin. *Nucleic Acids Res* **21**, 5157-5166 (1993).

42. Y. Pommier, Y. Sun, S. N. Huang, J. L. Nitiss, Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat Rev Mol Cell Biol* **17**, 703-721 (2016).

43. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

44. J. Ernst, M. Kellis, Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478-2492 (2017).

45. A. Nott *et al.*, Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134-1139 (2019).

46. D. Xing, L. Tan, C. H. Chang, H. Li, X. S. Xie, Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc Natl Acad Sci U S A* **118**, (2021).

47. S. M. Cerritelli, R. J. Crouch, The Balancing Act of Ribonucleotides in DNA. *Trends Biochem Sci* **41**, 434-445 (2016).

48. E. Crespan *et al.*, Impact of ribonucleotide incorporation by DNA polymerases beta and lambda on oxidative base excision repair. *Nat Commun* **7**, 10805 (2016).

49. Y. Pommier, A. Nussenzweig, S. Takeda, C. Austin, Human topoisomerases and their roles in genome stability and organization. *Nat Rev Mol Cell Biol* **23**, 407-427 (2022).

50. I. F. King *et al.*, Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501**, 58-62 (2013).

51. M. J. Zylka, J. M. Simon, B. D. Philpot, Gene length matters in neurons. *Neuron* **86**, 353-355 (2015).

52. G. Fragola *et al.*, Deletion of Topoisomerase 1 in excitatory neurons causes genomic instability and early onset neurodegeneration. *Nat Commun* **11**, 1962 (2020).

53. P. Pourquier *et al.*, Induction of reversible complexes between eukaryotic DNA topoisomerase I and DNA-containing oxidative base damages. 7, 8-dihydro-8-oxoguanine and 5-hydroxycytosine. *J Biol Chem* **274**, 8516-8523 (1999).

54. Y. J. Crow *et al.*, Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutieres syndrome and mimic congenital viral brain infection. *Nat Genet* **38**, 910-916 (2006).

55. Aditi *et al.*, Genome instability independent of type I interferon signaling drives neuropathology caused by impaired ribonucleotide excision repair. *Neuron* **109**, 3962-3979 e3966 (2021).

56. M. E. McCauley *et al.*, C9orf72 in myeloid cells suppresses STING-induced inflammation. *Nature* **585**, 96-101 (2020).

57. C. H. Yu *et al.*, TDP-43 Triggers Mitochondrial DNA Release via mPTP to Activate cGAS/STING in ALS. *Cell* **183**, 636-649 e618 (2020).

58. X. Xie *et al.*, Activation of innate immune cGAS-STING pathway contributes to Alzheimer's pathogenesis in 5xFAD mice. *Nat Aging* **3**, 202-212 (2023).

59. C. Marques *et al.*, Neuronal STING activation in amyotrophic lateral sclerosis and frontotemporal dementia. *Acta Neuropathol* **147**, 56 (2024).

60. G. D. Evrony *et al.*, Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496 (2012).

61. G. T. Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

62. K. C. Cotto *et al.*, Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat Commun* **14**, 1589 (2023).

63. Y. I. Li *et al.*, Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* **50**, 151-158 (2018).

64. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, (2013).

65. G. X. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).

66. S. J. Fleming *et al.*, Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat Methods* **20**, 1323-1335 (2023).

67. P. L. Germain, A. Lun, C. Garcia Meixide, W. Macnair, M. D. Robinson, Doublet identification in single-cell sequencing data using scDblFinder. *F1000Res* **10**, 979 (2021).

68. W. Muskovic, J. E. Powell, DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol* **22**, 329 (2021).

69. Y. Hao *et al.*, Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293-304 (2024).

70. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).

71. T. E. Bakken *et al.*, Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119 (2021).

72. J. Lai *et al.*, ATM-deficiency-induced microglial activation promotes neurodegeneration in ataxia-telangiectasia. *Cell Rep* **43**, 113622 (2024).

73. J. G. Harrison, W. J. Calder, V. Shastry, C. A. Buerkle, Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol Ecol Resour* **20**, 481-497 (2020).

74. L. He *et al.*, NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun Biol* **4**, 629 (2021).

75. M. I. Gabitto *et al.*, Integrated multimodal cell atlas of Alzheimer's disease. *Nat Neurosci* **27**, 2366-2383 (2024).

Acknowledgments:

We thank the Massachusetts Alzheimer's Disease Research Center and NIH NeuroBioBank for providing fresh frozen human tissues. We thank the donors and families for their contributions, and J. E. Neil for assistance with tissue procurement. We thank the Research Computing group at Harvard Medical School and Boston Children's Hospital.

Funding:

U.S. Department of Defense PRMRP Discovery Award W81XWH2010028 (Z.Z.)

Edward R. and Anne G. Lefler Center postdoctoral fellowship (Z.Z.)

American Heart Association Career Development Award 23CDA1046074 (Z.Z.)

National Institutes of Health grant DP2 AG086138 (M.B.M.)

National Institutes of Health grant R01 AG082346 (M.B.M.)

National Institutes of Health grant R56 AG079857 (A.Y.H.)

National Institutes of Health grant R01 AG088082 (A.Y.H.)

Alzheimer's Association Research Fellowship (A.Y.H.)

National Institutes of Health grant R01 HG012573 (P.J.P.)

National Institutes of Health grant K01 AG051791 (E.A.L.)

Suh Kyungbae Foundation (E.A.L.)

National Institutes of Health grant DP2 AG072437 (E.A.L.)

National Institutes of Health grant R01 NS032457 (C.A.W.)

National Institutes of Health grant R01 AG070921 (C.A.W., E.A.L.)

Massachusetts Alzheimer's Disease Research Center pilot grant (C.L.-T., C.A.W.)

Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation (C.A.W., E.A.L.)

Araminta Broch-Healey Endowed Chair in ALS (C.L.-T.)

Howard Hughes Medical Institute (C.A.W.)

Author contributions:

Conceptualization: Z.Z., L.J.L., G.D., P.J.P., C.L.-T., E.A.L., C.A.W.

Methodology: Z.Z., L.J.L., G.D., D.S., W.J.N., A.N.

Resource: M.B.M., A.Y.H., C.L.-T.

Software: L.J.L., G.D.

Investigation: Z.Z., L.J.L., G.D., Junho K., Jayoung K., K.K., B.S.

Visualization: Z.Z., L.J.L., G.D., Jayoung K.

Funding acquisition: Z.Z., P.J.P., C.L.-T., E.A.L., C.A.W.

Project administration: Z.Z., P.J.P., C.L.-T., E.A.L., C.A.W.

Supervision: P.J.P., C.L.-T., E.A.L., C.A.W.

Writing – original draft: Z.Z., L.J.L., G.D.

Writing – review & editing: Z.Z., L.J.L., G.D., Junho K., D.S., B.S., M.B.M., A.Y.H., P.J.P., C.L.-T., E.A.L., C.A.W.

Competing interests: P.J.P. is a member of the scientific advisory board for Bioskryb Genomics, Inc. C.L.-T serves on the scientific advisory board of SOLA Biosciences, Libra Therapeutics, Arbor Biotechnologies and Dewpoint Therapeutics and has received consultant fees from Mitsubishi Tanabe Pharma Holdings America, Sanofi and Applied Genetic Technologies Corporation. E.A.L. serves on the scientific advisory board of Genome Insight. C.A.W. is a paid consultant (cash, no equity) to Third Rock Ventures and Flagship Pioneering (cash, no equity) and is on the Clinical Advisory Board (cash and equity) of Maze Therapeutics. No research support is received. These companies did not fund and had no role in the conception or performance of this research project.

Data and materials availability: Previously published sequencing data were downloaded from dbGaP and NIAGADs. Sequencing data generated in this study will be deposited in a public repository, with controlled use conditions set by human privacy regulations. All scripts and code will be made publicly available on Zenodo and GitHub.

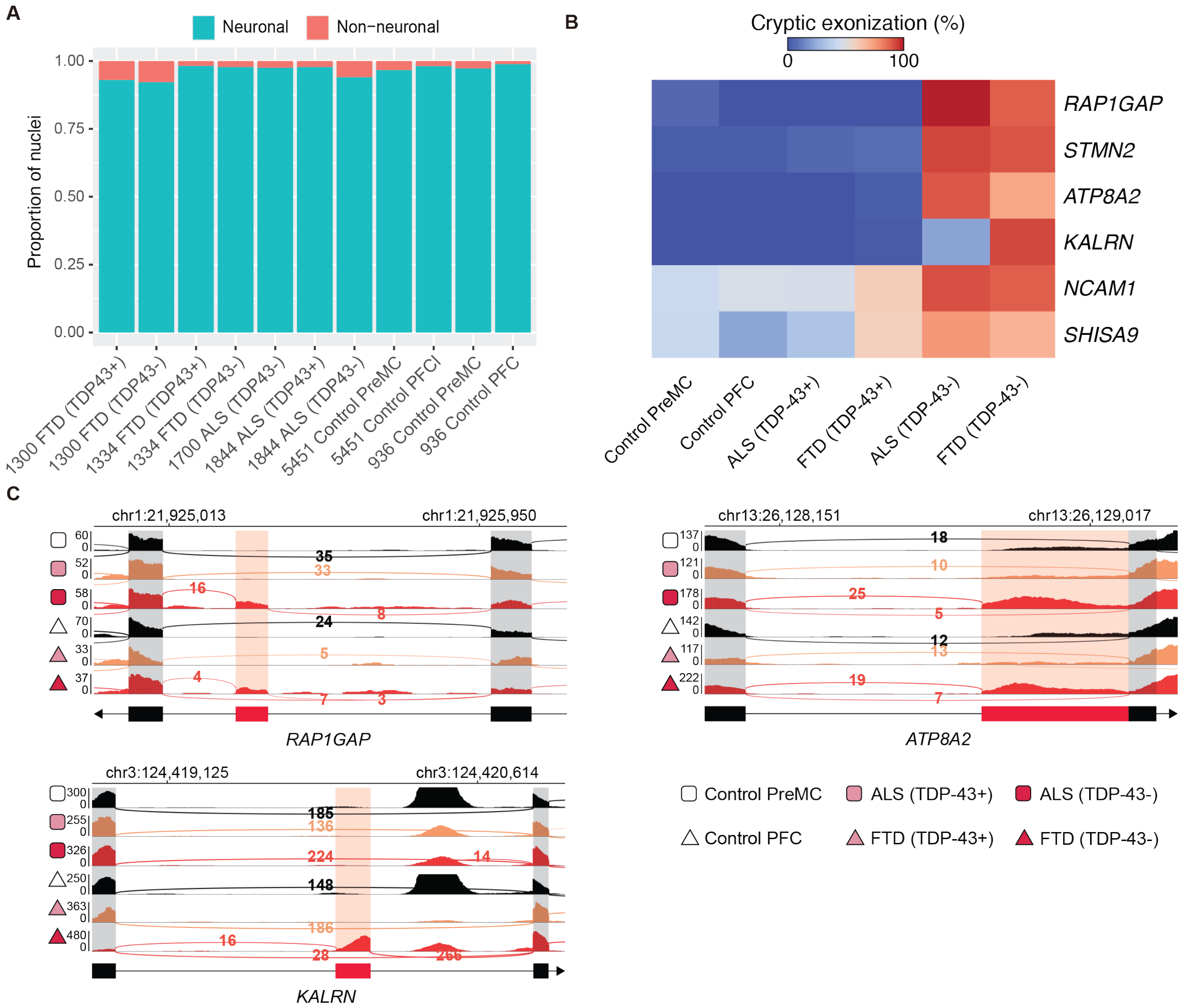

Fig. S1. Characterization of isolated TDP-43+ and TDP-43- neuronal nuclei from *C9ORF72* ALS and FTD brains. (A) Proportion of neuronal and non-neuronal nuclei across different cell and case conditions. Bars represent the fraction of neuronal (blue) and non-neuronal (red) nuclei for each dataset. (B) Heatmap showing the fraction of transcripts containing cryptic exons across various genes in different cell and case conditions. (C) IGV Genome browser tracks of *RAP1GAP*, *ATP8A2* and *KALRN* with snRNA-seq read coverage and splice junctions. TDP-43- neurons (red) show an increase in cryptic exon inclusion (highlighted region), which is absent or low in TDP-43+ (orange) and control (black) neurons. Numbers adjacent to splice junctions represent read counts supporting the junction.

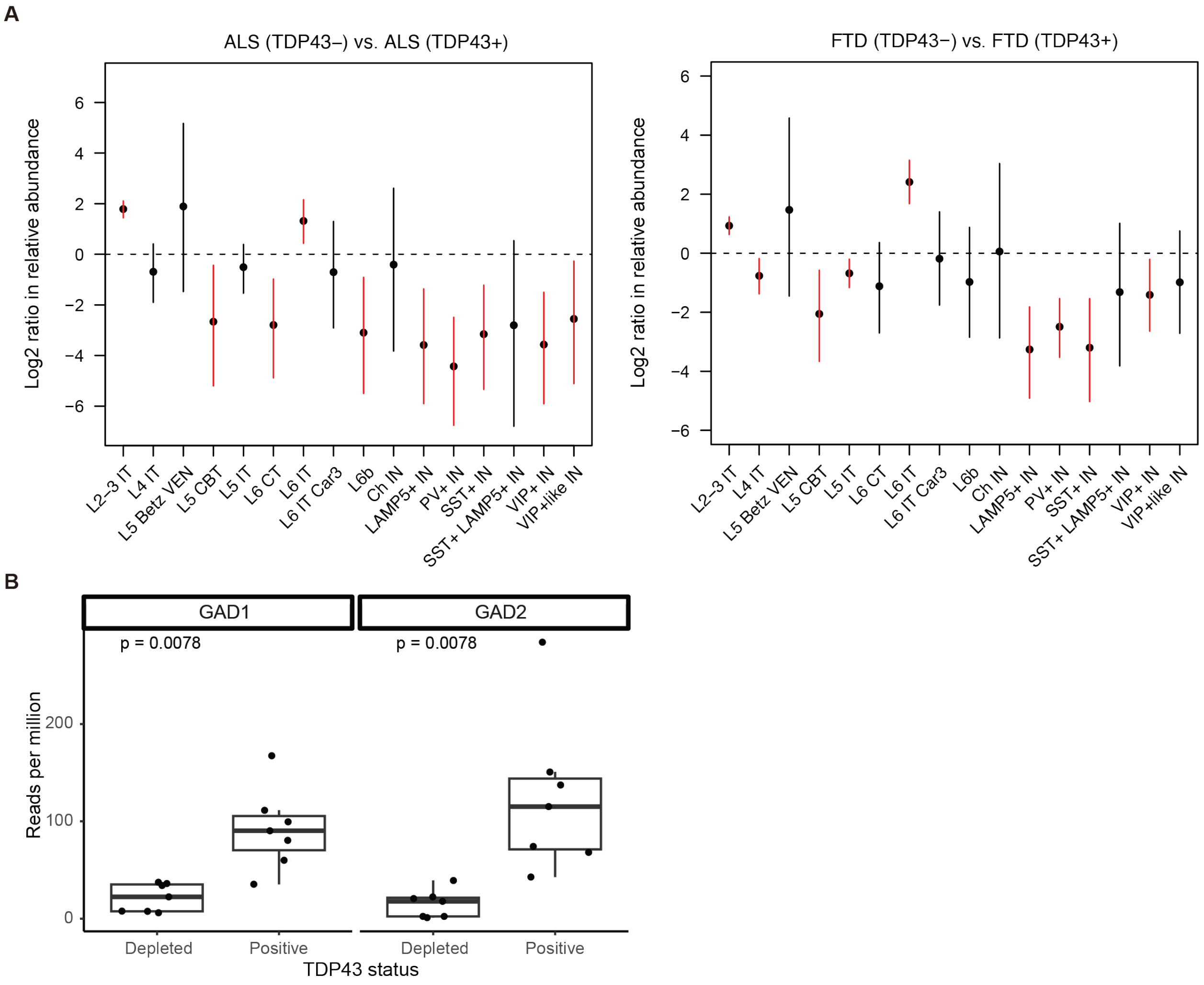

Fig. S2. Reduced inhibitory neuron abundance and marker gene expression in TDP-43- neurons from *C9ORF72* ALS and FTD brains. (A) Log_2 fold change in neuronal subtype abundance in TDP-43- vs. TDP-43+ neurons. Error bars show 95% credible intervals of posterior distributions from a Dirichlet multinomial model (see Methods). Red intervals indicate significant changes in neuronal subtype abundance. IT: intratelencephalic neurons. VEN: Von Economo neurons. CBT: corticobulbar tract neurons. CT: corticothalamic neurons. PV: parvalbumin. IN: inhibitory neurons. Ch IN: cholinergic inhibitory neurons. (B) Expression of inhibitory neuron marker genes in TDP-43- neurons. Box plots compare reads per million for *GAD1* and *GAD2* expression between TDP-43- and TDP-43+ neurons from *C9ORF72* FTD brains using previously generated bulk RNA-seq data. P-values are one-tailed paired Wilcoxon rank-sum tests.

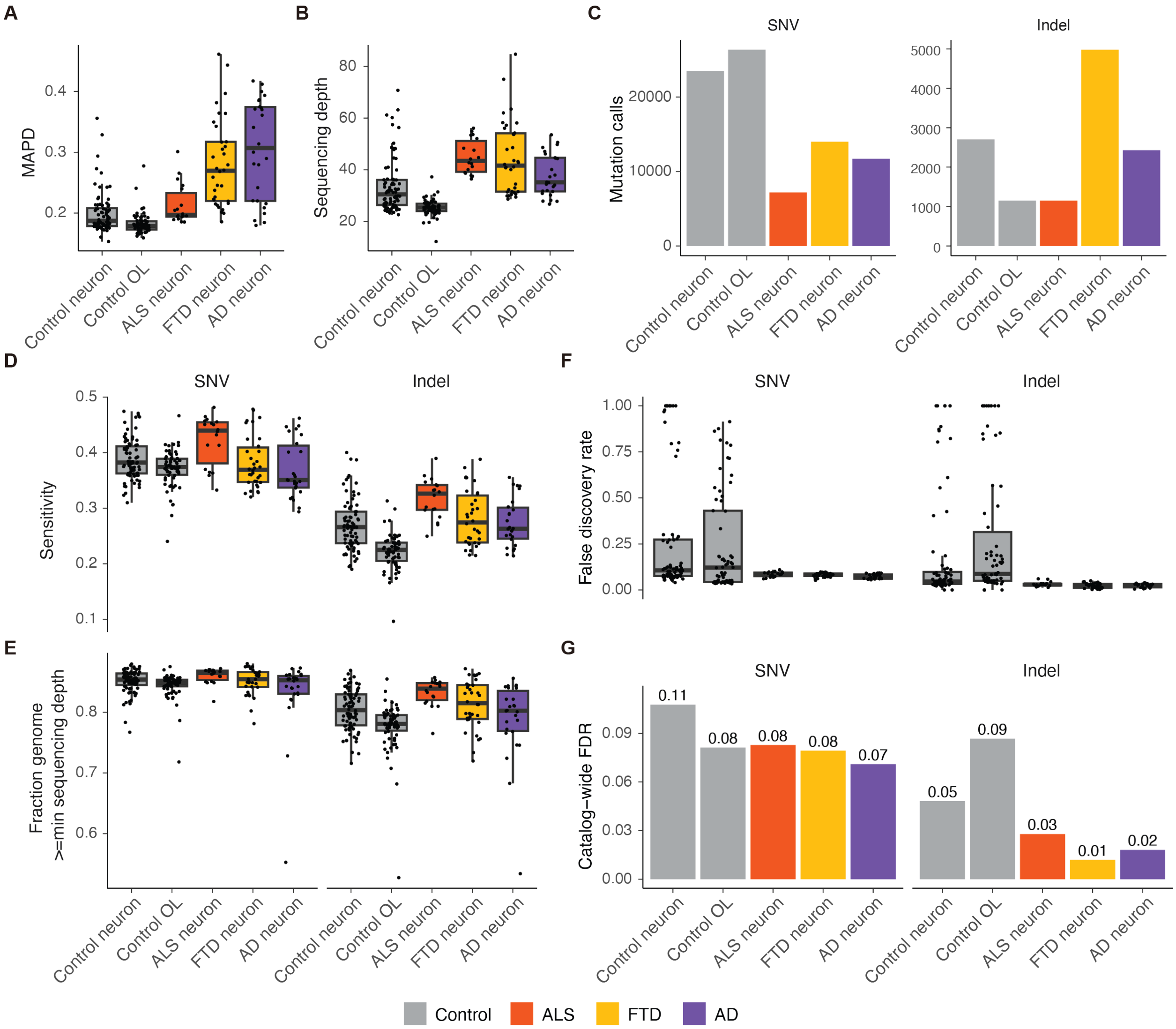

Fig. S3. scWGS metrics and quality assessment for PTA-amplified single cells. (A) Median Absolute Pairwise Difference (MAPD), a measure of amplification uniformity; low values indicate more uniform amplification. (B) Mean sequencing depth per cell. (C) Total numbers of somatic mutations called by SCAN2. (D) SCAN2 mutation detection sensitivity. (E) Fraction of the genome that passes the minimum read depth requirement for SCAN2. Different minimum read depths are used for sSNVs and sIndels. (F) Estimated false discovery rate per cell based on a previously published false positive rate from data simulations(*25*). (G) Estimated false discovery rate for each mutation catalog. That is, the sum of estimated false positives used in (F) divided by the total catalog sizes in (C). In all panels, points represent single cells. OL: oligodendrocyte.

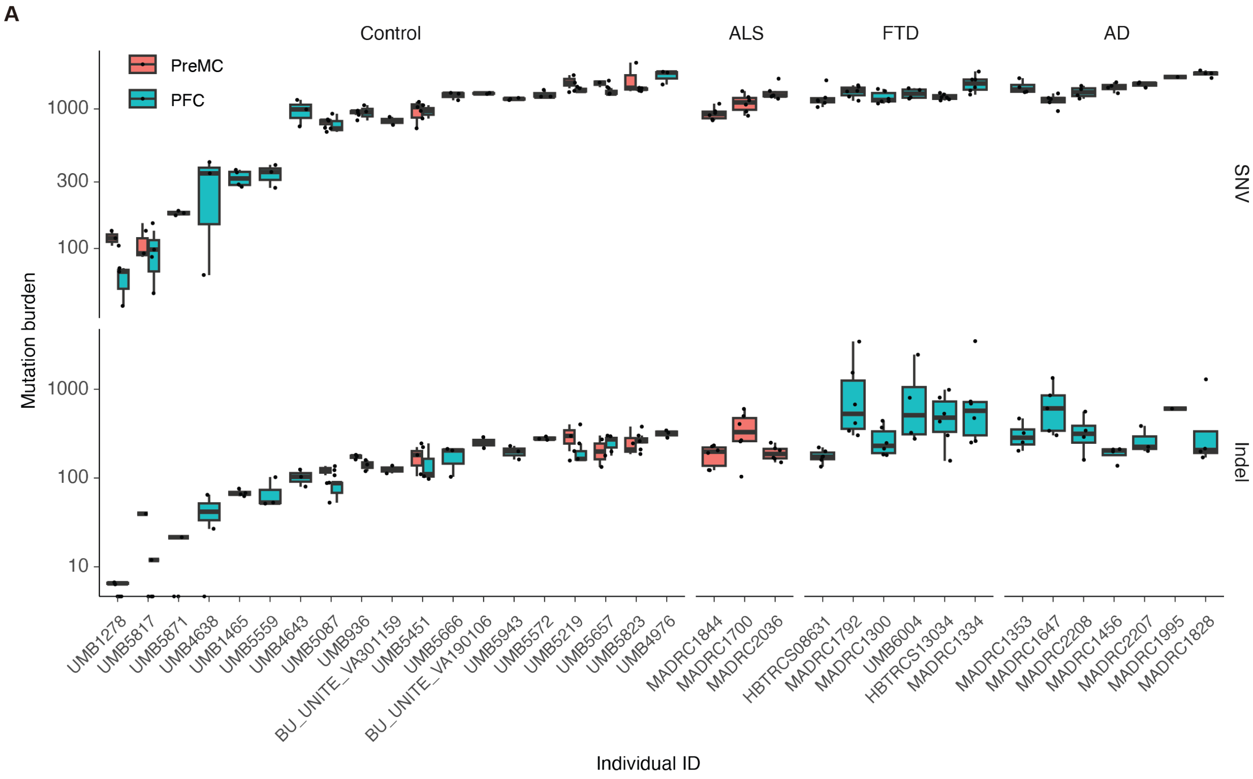

Fig. S4. Somatic Indel mutation burdens show large intra-individual variance. (A) Burdens of sSNVs and sIndels in neurons grouped by individual case. Red indicates PreMC (premotor cortex) neurons, while blue indicates PFC (prefrontal cortex) neurons. Mutation burdens are plotted on a logarithmic scale.

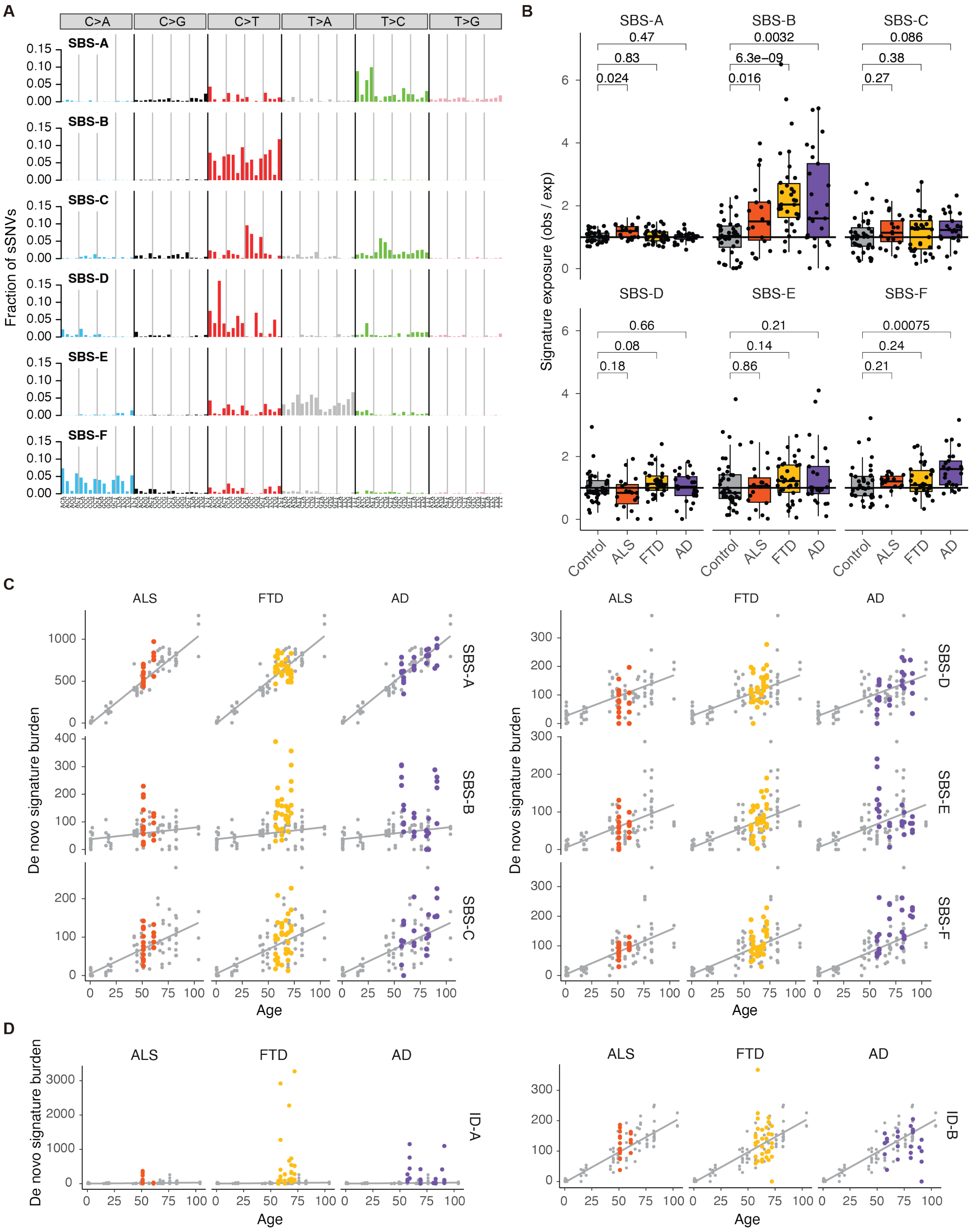

Fig. S5. *De novo* mutational signature analysis in PTA-amplified neurons. (A) Mutational spectra of *de novo* SBS signatures identified from joint analysis of diseased and neurotypical

control neurons amplified by PTA as well as previously published neurons and oligodendrocytes amplified using an older amplification technology (MDA). The x-axis shows the 96 possible trinucleotide mutation contexts, grouped by mutation type. The y-axis represents the fraction of sSNVs attributed to each signature. (B) Observed SBS signature exposure divided by the age-adjusted expected SBS signature exposure derived from control neurons (see Methods, and panels C-D). For control neurons, only observed/expected from individuals with age > 50 were included for statistical testing by Wilcoxon rank-sum test. (C-D) Burdens of *de novo* SBS and ID signatures in neurons vs. age. Points and regression lines in gray indicate the age-specific expected SBS and ID exposure (mixed-effects linear regression, Methods) and are identical in each row.

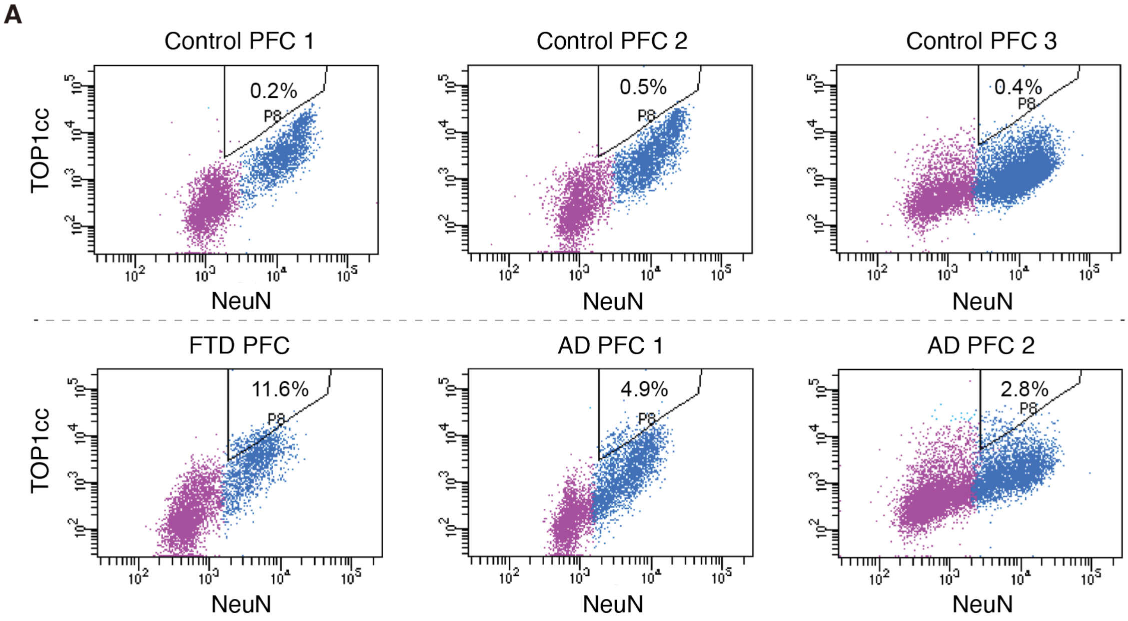

Fig. S6. Increased TOP1cc signals in neurons from *C9ORF72* FTD and AD brains compared to controls. (A) Flow cytometry analysis of TOP1cc levels in NeuN+ neurons from *C9ORF72* FTD, AD and control brains. Gates mark the population of neurons with elevated TOP1cc levels, with the percentage of TOP1cc+ neurons indicated in each plot. PFC: prefrontal cortex.

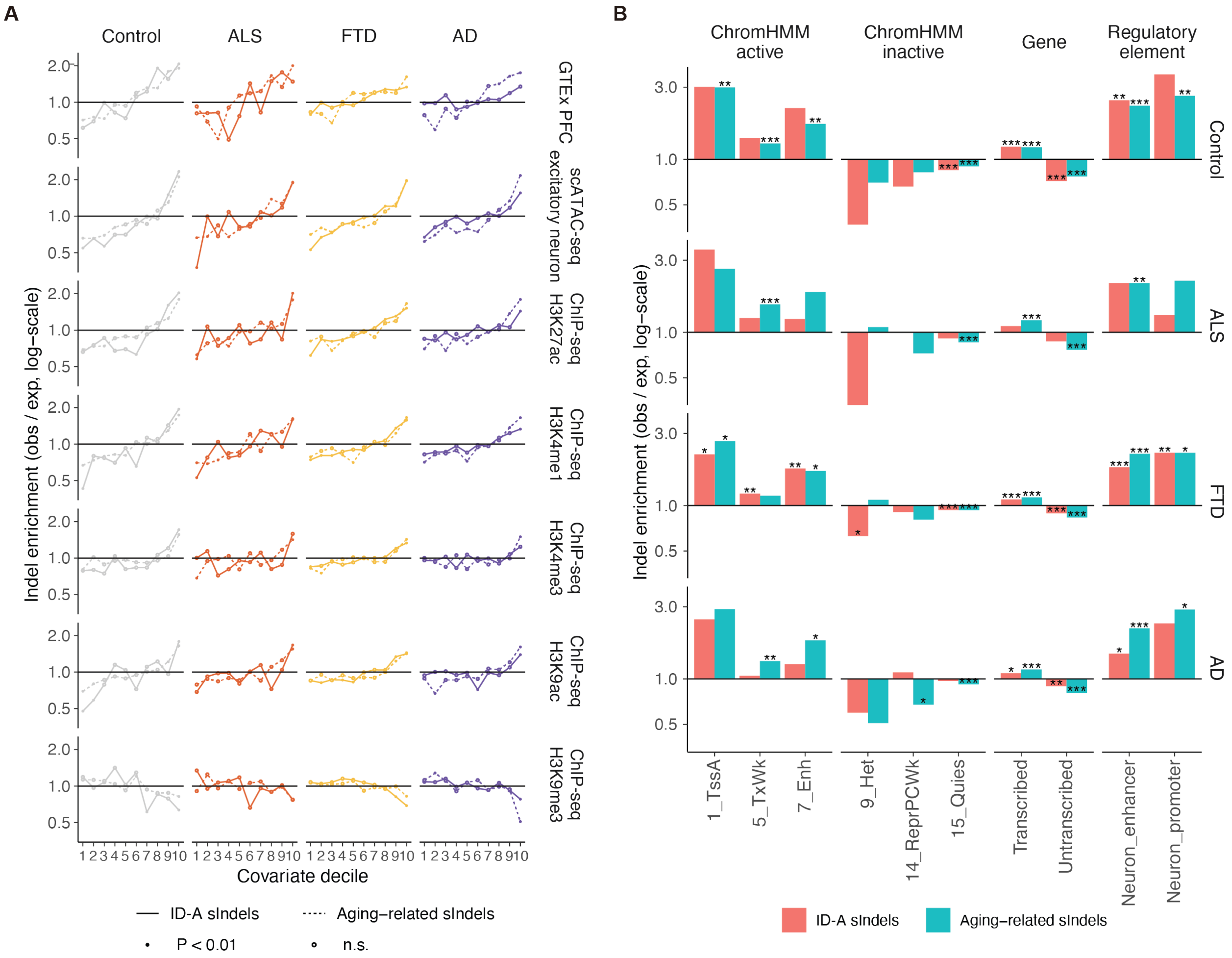

Fig. S7. ID-A-like indels are associated, though more weakly than aging-related sIndels, with transcriptionally active genomic regions. (A) sIndel enrichment analyses comparing sIndel burden with GTEx gene expression from prefrontal cortex; local chromatin accessibility in snATAC-seq data from neurotypical excitatory neurons; and histone modifications typical active and inactive chromatin from prefrontal cortex. Solid lines: ID-A-like sIndels, dotted lines: sIndels associated with normal aging. (B) Enrichment analyses comparing sIndel burdens with chromatin states in prefrontal cortex (determined by ChromHMM(*44*)), transcribed and untranscribed genomic regions defined by GENCODE v26, and promoters and enhancer elements identified as active in neurons(*45*). Red bars: ID-A-like sIndels, blue bars: sIndels associated with normal aging. Asterisks represent enrichment test P-values (Methods): * - p < 0.01, ** - p < 0.001, *** p < 0.0001. See Fig. 3A for the definition of ID-A-like sIndels.

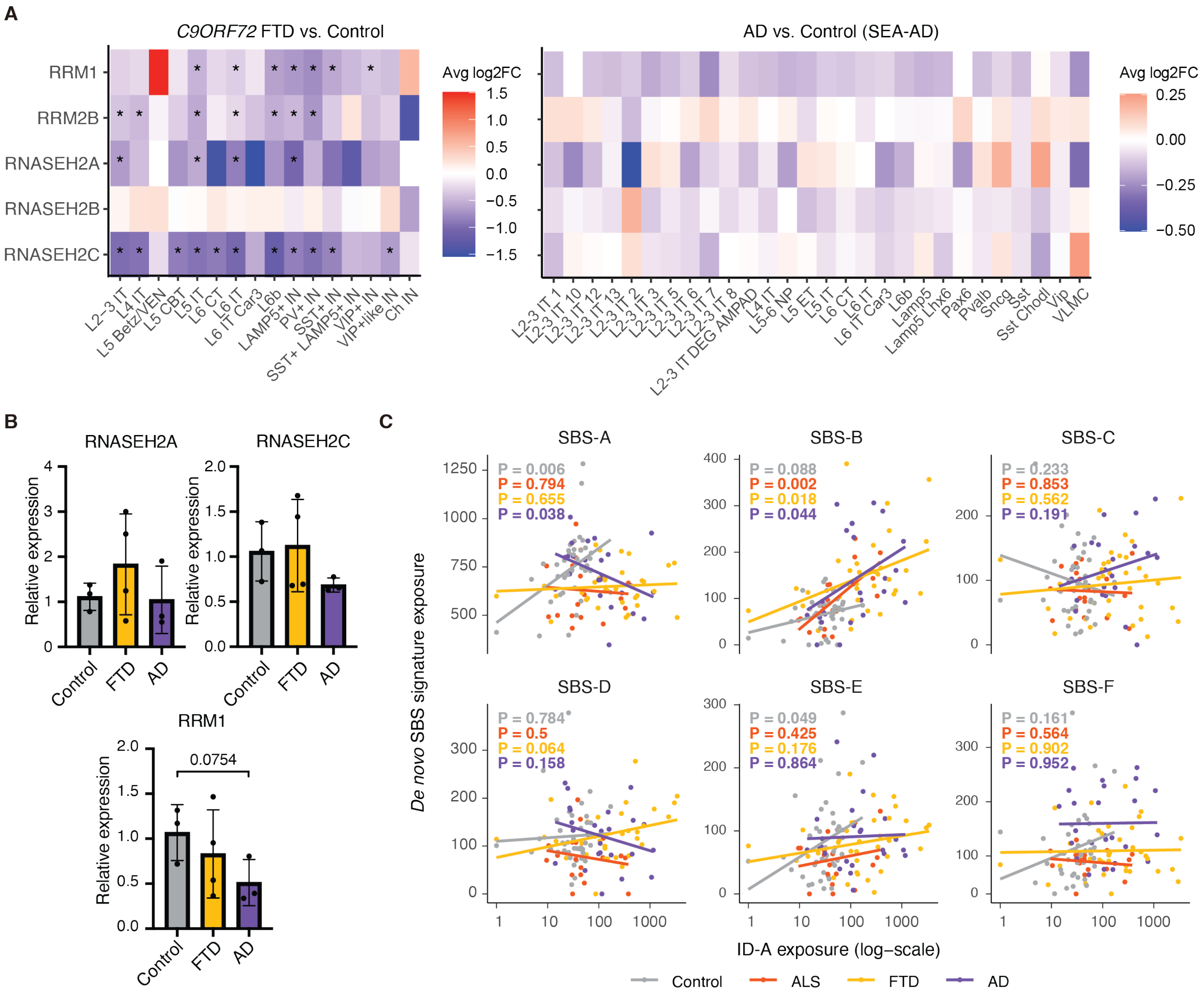

Fig. S8. Dysregulation of RNASEH2 and RRM genes and correlation between ID-A and oxidative stress in neurodegenerative conditions. (A) Heatmaps showing differential expression of RNASEH2 and RRM genes in neurons from *C9ORF72* FTD and AD brains compared to controls. The AD vs. control comparisons is based on the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) dataset. Color scale represents average $\log_2$FC, with red indicating upregulation and blue indicating downregulation. FDR-adjusted P-values < 0.05 are marked by "*". IT: intratelencephalic neurons. VEN: Von Economo neurons. CBT: corticobulbar tract neurons. CT: corticothalamic neurons. PV: parvalbumin. IN: inhibitory neurons. Ch IN: cholinergic inhibitory neurons. NP: near-projecting neurons. ET: extratelencephalic neurons. VLMC: vascular leptomeningeal cells. (B) Expression levels of *RNASEH2A, RNASEH2C* and *RRM1* in neurons measured by qPCR. P-value is two-tailed unpaired *t*-test. (C) Correlation

between ID-A and all six *de novo* SBS signatures. P-values are regression t-tests on slope, and text color indicates phenotype.

Table S1.

Information of the control and neurodegenerative diseases cases and samples.

donor	age	sex	phenotype	*C9ORF72* repeat expansion
UMB1465	17.5	M	Control	
UMB4638	15.1	F	Control	
UMB4643	42.2	F	Control	
UMB1278	0.4	M	Control	
UMB5171	79.2	M	Control	
UMB5219	76	F	Control	
UMB5817	0.6	M	Control	
UMB5511	80.2	F	Control	
UMB5823	82.7	F	Control	
UMB5840	75.3	M	Control	
UMB936	49.2	F	Control	
UMB5559	19.8	F	Control	
UMB5657	82.2	M	Control	
UMB5087	44.9	M	Control	
UMB5532	18.4	M	Control	
UMB5871	2	M	Control	
MADRC1353	57	F	Alzheimers	
MADRC1456	81	M	Alzheimers	
MADRC1647	59	F	Alzheimers	
MADRC1828	91	F	Alzheimers	
MADRC1995	89	F	Alzheimers	
MADRC2208	69	F	Alzheimers	
MADRC4556	70	F	Alzheimers	
MADRC5222	80	F	Alzheimers	
UMB4976	104	F	Control	
UMB5451	57	F	Control	
UMB5943	69	M	Control	
BU_UNITE_VA190106	67	M	Control	
BU_UNITE_VA301159	51	M	Control	
UMB5572	70	F	Control	
UMB5666	65	M	Control	
MADRC2207	83	M	Alzheimers	
MADRC2036	61	M	ALS	Yes
MADRC1844	51	F	ALS	Yes
MADRC1700	51	F	ALS	Yes
MADRC1300	63	M	FTD	Yes
MADRC1334	72	M	FTD	Yes
MADRC1792	59	F	FTD	Yes
UMB6004	67	M	FTD	Yes
HBTRCS08631	57	F	FTD	Yes
HBTRCS13034	69	F	FTD	Yes

Table S2.

Number of cells passing and failing QC across cell types and conditions.

amp	phenotype	celltype	QC	Mean age	Individuals	N male cells	Cells	TDP43- cells	PFC cells	BA6 cells	DG cells
MDA	Control	neuron	PASS	51.6	18	116	190	0	158	6	26
MDA	Control	oligo	PASS	41	4	20	20	0	20	0	0
MDA	Control	mixedglia	PASS	41	4	20	20	0	20	0	0
MDA	AD	neuron	PASS	74.4	8	9	81	0	81	0	0
MDA	AD	neuron	FAIL	84.7	3	0	6	0	6	0	0
MDA	Control	neuron	FAIL	79.2	1	1	1	0	1	0	0
PTA	Control	neuron	PASS	47.8	19	41	80	0	56	24	0
PTA	Control	oligo	PASS	38.2	12	37	66	0	66	0	0
PTA	AD	neuron	PASS	72.5	7	7	26	0	26	0	0
PTA	AD	neuron	FAIL	89	1	0	3	0	3	0	0
PTA	ALS	neuron	PASS	54.3	3	6	18	9	0	18	0
PTA	FTD	neuron	PASS	64.4	6	16	34	18	34	0	0
PTA	FTD	neuron	FAIL	67	1	2	2	0	2	0	0