RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Diverse somatic genomic alterations in single neurons in chronic traumatic encephalopathy

Guanlan Dong†, Chanthia C. Ma†, Shulin Mao‡, Katherine Sun-Mi Brown‡, Samuel M. Naik, Gannon A. McDonough, Samadhi P. Wijethunga, Junho Kim, Samantha L. Kirkham, Diane D. Shao, Jonathan D. Cherry, Madeline Uretsky, Elizabeth Spurlock, Ann C. McKee, August Yue Huang*, Michael B. Miller*, Eunjung Alice Lee*, Christopher A. Walsh*

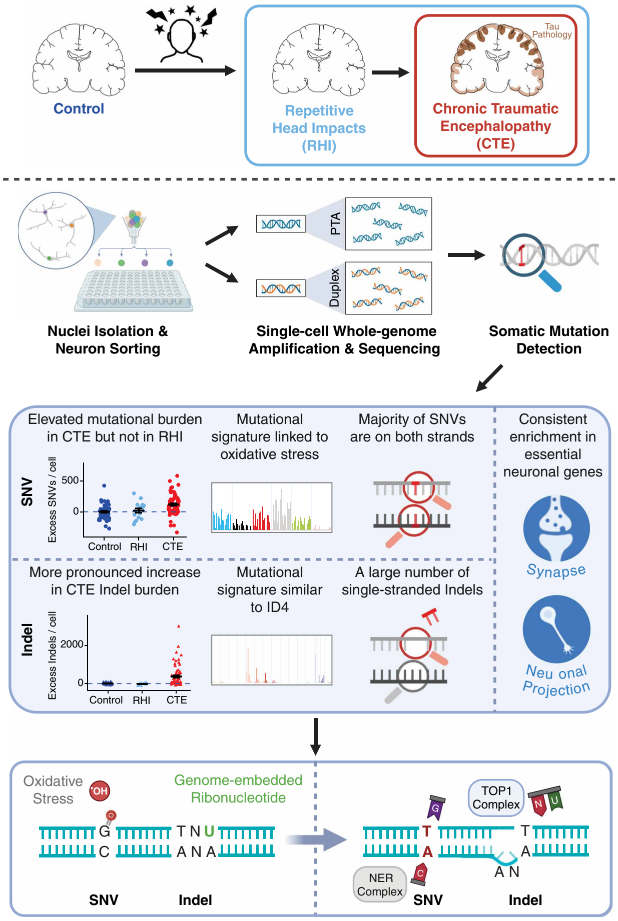
Full article and list of author affiliations: https://doi.org/10.1126/science.adu1351

INTRODUCTION: Chronic traumatic encephalopathy (CTE) is a neurodegenerative disease that develops years after exposure to repetitive head impacts (RHI) in some individuals. It is most often observed in individuals with a history of playing contact sports. Clinically, CTE is diagnosed through postmortem neuropathological examination of brain tau deposition, a hallmark it shares with Alzheimer's disease (AD). However, CTE is distinct in its tau accumulation patterns and clinical symptoms. Moreover, not everyone exposed to RHI develops CTE, and little is known about how head trauma can lead to CTE pathogenesis.

RATIONALE: While somatic mutations are well established as drivers for cancer progression, they also accumulate in noncancer cells and even in nondividing cells such as neurons. Recent advances in single-cell whole-genome sequencing (scWGS) technology have enabled the detection of private somatic mutations that accumulate in each nondividing neuron. Previous studies of single neurons showed that private somatic single-nucleotide variants (SNVs) increase during normal aging and more so in neurodegenerative diseases. However, the profiles of other mutation types have not been investigated in disease, and the role of somatic mutations in CTE is unknown.

RESULTS: Here we show that somatic mutations significantly increase in neurons of CTE patients and present specific disease-related patterns. We applied two scWGS methods to hundreds of neurons isolated from the prefrontal cortex of 15 individuals with CTE and 4 individuals with RHI but without CTE and compared mutational burdens and signatures with neurons from neurotypical controls and AD patients. In addition to the conventional strand-agnostic method, we included a strand-aware duplex sequencing modality that can distinguish between double-stranded mutations and single-stranded DNA lesions. We found a significant elevation of mostly double-stranded somatic SNVs in CTE that resembles a pattern previously reported in AD. Furthermore, we discovered high burdens of short deletions in both CTE and AD, resembling a known pattern of COSMIC signature ID4. Notably, the high burdens are only found in a subset of CTE and AD patients with a prominent signal from single-stranded events. Furthermore, somatic mutations in protein-coding genes are consistently enriched in essential neuronal functions, which, coupled with increased burdens in CTE, may lead to a dysregulated neuronal transcriptome and contribute to neurodegeneration. In contrast, somatic mutational burdens in individuals exposed to RHI but without a CTE diagnosis are similar to those of neurotypical controls, suggesting that the development of CTE involves mechanisms beyond those caused by RHI alone.

CONCLUSION: In this study, we characterized two types of somatic mutations in CTE neurons using both strand-agnostic and strand-specific methods of scWGS. Our results reveal that neurons in the CTE brain are exposed to stereotyped mutational processes shared with AD, implying potentially common pathogenic mechanisms. Specifically, increased SNVs resemble a pattern previously linked to oxidative damage, and increased short deletions resemble a pattern associated with topoisomerase 1 (TOP1) activity. In addition, the absence of similar changes in neurons exposed to RHI but without CTE highlights the complexity of CTE pathogenesis. Further study on the mechanisms leading to the short deletion pattern in certain CTE and AD patients may contribute to identifying novel therapeutic targets for neurodegenerative diseases. □


Control
Repetitive Head Impacts (RHI)
Tau Pathology
Chronic Traumatic Encephalopathy (CTE)
PTA
Duplex
Nuclei Isolation & Neuron Sorting
Single-cell Whole-genome Amplification & Sequencing
Somatic Mutation Detection
SNV
Elevated mutational burden in CTE but not in RHI
Excess SNVs / cell
500
0
Control
RHI
CTE
Mutational signature linked to oxidative stress
Majority of SNVs are on both strands
Indel
More pronounced increase in CTE Indel burden
Excess Indels / cell
2000
0
Control
RHI
CTE
Mutational signature similar to ID4
A large number of single-stranded Indels
Consistent enrichment in essential neuronal genes
Synapse
Neu onal Projection
Oxidative Stress
OH
Genome-embedded Ribonucleotide
G
C
T N U
A N A
SNV
Indel
TOP1 Complex
G
N
U
T
A
T
A
A N
NER Complex
C
SNV
Indel

Excess of somatic mutations in single neurons in CTE. CTE develops after exposure to RHI in some individuals. Somatic mutations were profiled from single neurons in CTE brains using two scWGS methods. An excess of somatic SNVs and short deletions was found in CTE, resembling patterns associated with distinct mutagenic mechanisms that may contribute to CTE pathogenesis.

*Corresponding author. Email: christopher.walsh@childrens.harvard.edu (C.A.W.); ealice.lee@childrens.harvard.edu (E.A.L.); mbmiller@bwh.harvard.edu (M.B.M.); yue.huang@childrens.harvard.edu (A.Y.H.) †These authors contributed equally to this work. ‡These authors contributed equally to this work. Cite this article as G. Dong *et al.*, *Science* 390, eadu1351 (2025). DOI: 10.1126/science.adu1351

NEUROSCIENCE

Diverse somatic genomic alterations in single neurons in chronic traumatic encephalopathy

Guanlan Dong[1,2,3]†, Chanthia C. Ma[1,2,4,5]†, Shulin Mao[1,2,6]‡, Katherine Sun-Mi Brown[5]‡, Samuel M. Naik[5], Gannon A. McDonough[5], Samadhi P. Wijethunga[5], Junho Kim[7], Samantha L. Kirkham[1,2], Diane D. Shao[1,2,8], Jonathan D. Cherry[9,10,11], Madeline Uretsky[10], Elizabeth Spurlock[10], Ann C. McKee[9,10,11,12], August Yue Huang[1,2,13]*, Michael B. Miller[1,2,5,13]*, Eunjung Alice Lee[1,2,13]*, Christopher A. Walsh[1,2,13,14]*

Chronic traumatic encephalopathy (CTE) is a neurodegenerative disease linked to exposure to repetitive head impacts (RHI), yet little is known about its pathogenesis. Applying two single-cell whole-genome sequencing methods to hundreds of neurons from prefrontal cortex of 15 individuals with CTE and 4 with RHI without CTE, we revealed increased somatic single-nucleotide variants in CTE, exhibiting a pattern previously reported in Alzheimer's disease (AD). Furthermore, we discovered high burdens of somatic small insertions and deletions in a subset of CTE individuals, resembling a known pattern, ID4, also found in AD. Our results suggest that neurons in CTE experience stereotyped mutational processes shared with AD; the absence of similar changes in RHI neurons without CTE suggests that CTE involves mechanisms beyond RHI alone.

Chronic traumatic encephalopathy (CTE) develops years after exposure to repetitive head impacts (RHI) and is most often found in athletes playing contact sports (*1*–*5*). CTE is diagnosed by postmortem neuropathological examination, on the basis of the presence of a pathognomonic lesion consisting of perivascular hyperphosphorylated tau neurofibrillary tangles in neurons at the depths of the cortical sulci (*1*, *2*). Tau deposition is common to both CTE and Alzheimer's disease (AD) (*4*, *6*); however, CTE is characterized by distinct neuropathological features, tau molecular structural conformation (*7*, *8*), and clinical symptoms (*5*, *9*). The precise mechanisms by which RHI induce tau neurodegeneration are poorly understood. Somatic mutations are well-known drivers of cellular proliferation in neoplasia (*10*–*12*), but they also accumulate in nonneoplastic cells and even nondividing neurons with age (*13*–*17*). Multiple studies using single-cell whole-genome sequencing (scWGS) (*17*, *18*) and single-molecule duplex sequencing (*16*, *19*, *20*) converge on accumulation rates of 16 to 17 somatic single-nucleotide variants (sSNVs) per year and two to three somatic small insertions and deletions (sIndels) per year in neurons from neurotypical controls. Furthermore, neurons from individuals with AD and other neurodegenerative disorders show higher sSNV burden with distinct mutational signatures (*18*, *21*). In this study, we applied scWGS to amplified DNA from single neuronal nuclei isolated from CTE, RHI without CTE, AD, and neurotypical control individuals (Fig. 1A), using both strand-agnostic and strand-specific amplification methods to assess somatic mutations in CTE.

sSNVs in CTE neurons

We acquired frozen brain tissue from dorsolateral prefrontal cortex (PFC) from individuals with a history of RHI, with and without CTE, as well as neurotypical controls. We isolated nuclei of single neurons by staining for the nuclear dye 4′,6-diamidino-2-phenylindole (DAPI) and the pan-neuronal marker NeuN and sorted neurons using fluorescence-activated nuclei sorting (FANS), gating specifically for the largest NeuN-positive nuclei (Fig. 1B and fig. S1). This gating method isolates pyramidal excitatory neurons with a purity of >99% (*18*). We sorted single nuclei into individual wells of 96-well plates then performed whole-genome amplification using primary template amplification (PTA) (*22*) or multiplexed end-tagging amplification of complementary strands (META-CS) (*19*). Amplified DNA underwent multiple screening and quality control steps, ensuring that only well-amplified genomes were used for sequencing. We sequenced PTA-amplified genomes of 68 neurons from 15 cases of CTE (exposed to RHI with neuropathologically verified CTE) and 16 neurons from 4 cases of RHI (exposed to RHI without CTE) and compared these with 56 neurons from 19 neurotypical controls (*23*) and 27 neurons from 7 cases of AD (*18*) (tables S1 to S3).

We identified both sSNVs and sIndels in each single PTA neuron using SCAN2 (*17*) and found significantly increased sSNVs in CTE, averaging 114 more sSNVs per cell compared with controls [$P = 0.003$, linear mixed-effects (LME) model; Fig. 1C]. The increased sSNV burdens are comparable to those in AD (Fig. 1D). In contrast, sSNV numbers in RHI cases were indistinguishable from controls ($P = 0.725$, LME model; Fig. 1C), suggesting that increased sSNVs in CTE arise not from RHI exposure alone but from additional factors associated with the development of CTE. All CTE and RHI brain samples were collected and processed at the UNITE Brain Bank to control for potential batch effects, although the small sample size of RHI may limit statistical power. We performed nonparametric tests on somatic burdens after adjusting for age, where both CTE and AD neurons showed a significant excess of sSNVs compared with controls (CTE, $P = 3.3 \times 10^{-6}$; AD, $P = 2.3 \times 10^{-5}$; two-tailed Wilcoxon test; Fig. 1E and fig. S2, A to C). The increased sSNVs in CTE remained significant after controlling for additional quality metrics (see methods and fig. S3). Furthermore, although neurons in each case showed some variation, we observed greater mean sSNVs in nearly all CTE cases than the burden attributable to normal aging (Fig. 1F).

Somatic mutations were broadly distributed across each neuron's genome (Fig. 1G). Notably, we found no association between sSNV burden and potential risk factors for CTE pathology, including cortex topology (*1*) (fig. S4) and years of playing American football (*6*, *24*–*26*), which implies that the duration of RHI exposure does not directly lead to increased sSNVs, in line with the similar sSNV burden between RHI and controls.

To investigate the strand specificity of excess sSNVs in CTE, we used a modified duplex META-CS method (*19*) (see methods) to profile 115 neurons from 12 CTE cases and 146 neurons from 16 neurotypical controls with paired PTA data (table S1). We distinguished variants representing double-stranded mutations and single-stranded DNA lesions

[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]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA, USA. [4]Harvard-MIT MD-PhD Program, Harvard Medical School, Boston, MA, USA. [5]Division of Neuropathology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. [7]Department of Biological Sciences, Sungkyunkwan University, Suwon, South Korea. [8]Department of Neurology, Boston Children's Hospital, Boston, MA, USA. [9]Veterans Affairs (VA) Boston Healthcare System, US Department of Veteran Affairs, Boston, MA, USA. [10]Alzheimer's Disease Research Center and Chronic Traumatic Encephalopathy Center, Chobanian and Avedisian School of Medicine, Boston University, Boston, MA, USA. [11]Department of Pathology and Laboratory Medicine, Chobanian and Avedisian School of Medicine, Boston University, Boston, MA, USA. [12]Department of Neurology, Chobanian and Avedisian School of Medicine, Boston University, Boston, MA, USA. [13]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [14]Boston Children's Hospital, Howard Hughes Medical Institute, Boston, MA, USA. ***Corresponding author. Email: christopher.walsh@childrens.harvard.edu (C.A.W.); ealice.lee@childrens.harvard.edu (E.A.L.); mbmiller@bwh.harvard.edu (M.B.M.); yue.huang@childrens.harvard.edu (A.Y.H.)** †These authors contributed equally to this work. ‡These authors contributed equally to this work.

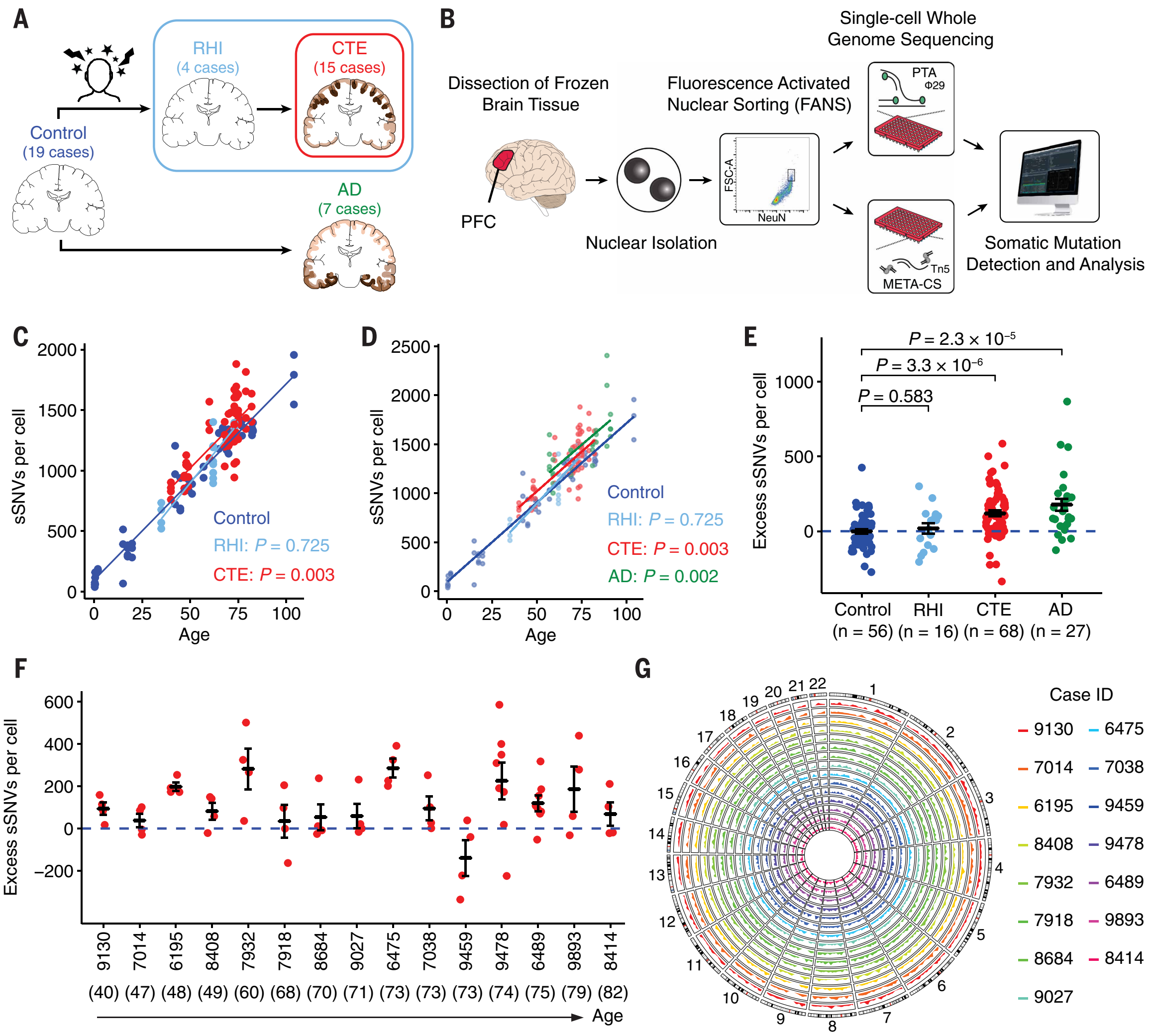


Fig. 1. Study design and sSNV burden in CTE neurons. (**A**) Cohort design: neurotypical control, repetitive head impacts (RHI), chronic traumatic encephalopathy (CTE), and Alzheimer's disease (AD). Illustrations of CTE and AD show characteristic tau pathology patterns (in brown). The number of cases included in this study is shown for each clinical condition. (**B**) Single-cell whole-genome sequencing (scWGS) experimental approach. Nuclei were isolated from postmortem human brain tissue and subjected to fluorescence-activated nuclear sorting for the neuronal nuclear marker NeuN. Nuclei were sorted, lysed, and subjected to primary template amplification (PTA) and multiplexed end-tagging amplification of complementary strands (META-CS). Amplified genomic DNA was then assayed using WGS to identify somatic mutations. PTA data were used to determine somatic mutation burden, while META-CS data were used to identify strand-related signatures. (**C**) sSNV burden in CTE (red), RHI (light blue), and neurotypical control (dark blue) brains with a significant increase in CTE but not in RHI when compared with controls. sSNV burden (from each neuron as a point) estimated by SCAN2 is fitted against age by clinical conditions using LME models (neurotypical control, dark blue; RHI, light blue, $P = 0.725$; CTE, red, $P = 0.003$). P values compare each clinical condition against controls. (**D**) Similar to (C), with added neurons from AD brains (green, $P = 0.002$ using the LME model). P values compare each clinical condition against controls. (**E**) Excess sSNV burden in RHI, CTE, and AD compared with neurotypical control after adjusting for age. Data are mean ± standard error. The dashed blue line shows sSNVs attributable to age (zero excess). P values are from two-tailed Wilcoxon tests. (**F**) Excess sSNVs in each CTE case ordered by increasing age (indicated in parentheses). The dashed blue line shows sSNVs attributable to age (zero excess). (**G**) Circos plot showing the PTA sSNV density distribution of CTE cases across the whole genome. Each CTE case is depicted by color in a circular track.

(fig. S5A). According to META-CS–derived strand-specific signatures, PTA-profiled sSNVs showed, on average, 85% double-stranded signature contribution (fig. S5B), indicating that PTA-identified sSNVs are predominantly double-stranded, albeit not exclusively. This aligns with the highly similar rate of age-related sSNV increase when profiled by either PTA ([17]) or strand-specific methods ([16], [19]).

Mutational signatures of sSNVs in CTE neurons

To investigate potential sources of sSNVs in CTE neurons, we performed mutational signature analysis on the PTA scWGS data from CTE and neurotypical control neurons. First, we decomposed the sSNVs into signatures A and C, previously identified in single neurons ([21]) (Fig. 2, A and B). Signature A closely resembles SBS5 from the COSMIC database (v.3.2) ([27]), a clocklike signature associated with aging (SBS, single base substitution). Signature C resembles SBS8, a signature associated with deficient transcription-coupled nucleotide excision repair (TC-NER) ([21], [28]), along with other SBS signatures, which have also been linked with oxidative damage in AD neurons ([18]). The contribution of signature A increased with age without showing differences between CTE and control neurons (Fig. 2, C and D). In contrast, we observed a significantly larger contribution of signature C in CTE neurons, with an average excess of 64 sSNVs per cell ($P = 7.7 \times 10^{-4}$, LME

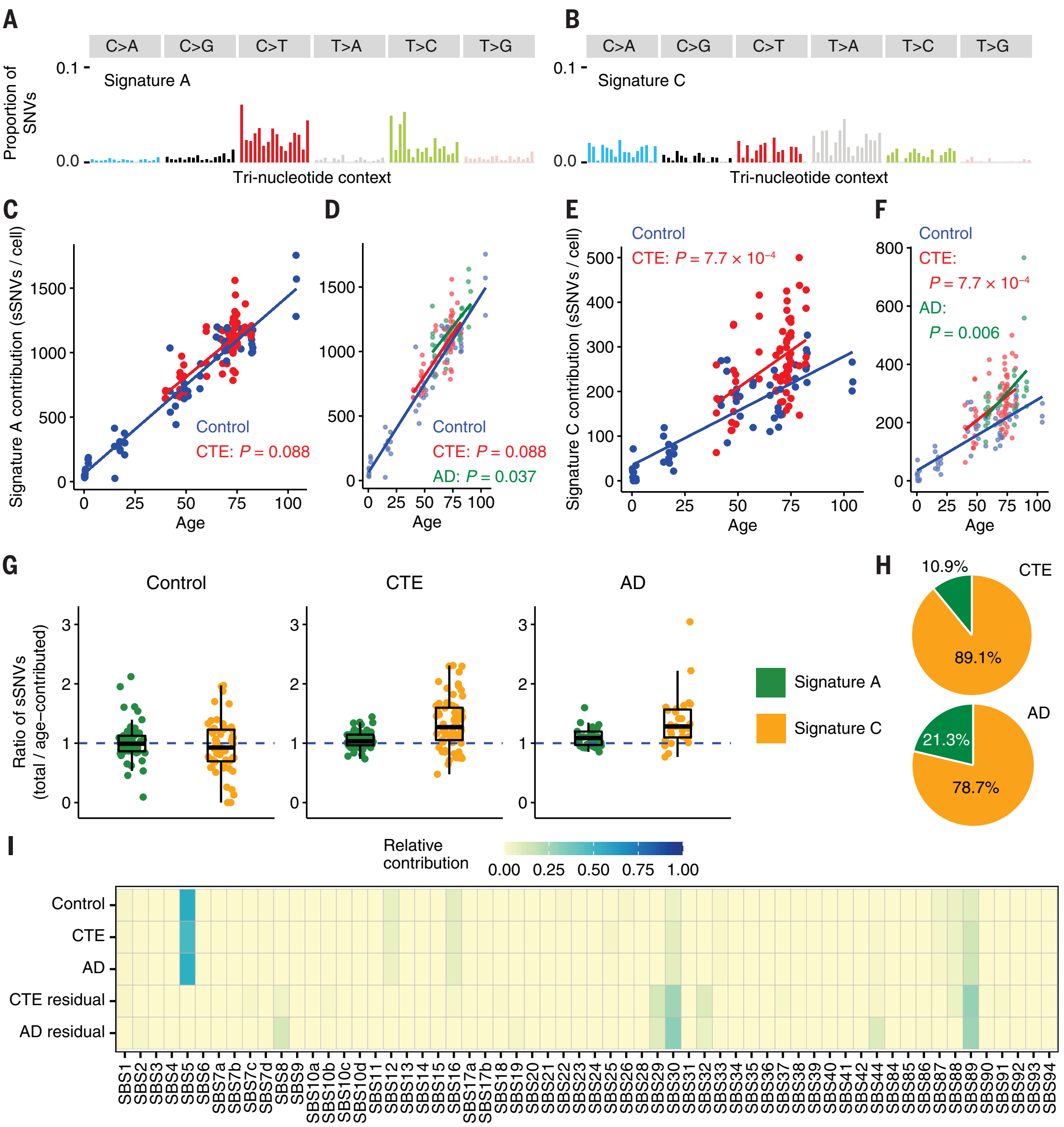


Fig. 2. sSNV mutational signatures in CTE neurons. (**A** and **B**) Mutation spectra of signatures A and C. (**C** to **F**) sSNV burden separated by contributions from signature A [(C) and (D)] and signature C [(E) and (F)] in CTE (red), AD (green), and neurotypical control (dark blue) neurons. These signatures decompose sSNV burden into age- and disease-specific effects. Their contributions are fitted against age by clinical conditions using LME models [(C), signature A, CTE: $P = 0.088$; (D), signature A, CTE: $P = 0.088$, AD: $P = 0.037$; (E), signature C, CTE: $P = 7.7 \times 10^{-4}$; (F), signature C, CTE: $P = 7.7 \times 10^{-4}$, AD: $P = 0.006$]. P values compare each clinical condition against controls. (**G**) Ratios of each signature's total contribution to age-related contribution in neurotypical control, CTE, and AD neurons. Age-contributed sSNVs of each signature are obtained from the LME model in neurotypical controls. A ratio of >1 indicates a higher contribution from the signature compared with age-matched controls. Bars in each box plot from top to bottom show the first, second (median), and third quartile; whiskers extend 1.5 interquartile range (IQR). (**H**) Relative contribution of signatures A and C in CTE and AD after adjusting for age. Based on the ratios shown in (G), the relative contribution of each signature is calculated by removing the median ratio of neurotypical controls (representing age effect) from the median ratio of each disease. (**I**) Mutational spectra of neurotypical control, CTE, and AD neurons are fitted to the COSMIC SBS database of cancer mutational signatures. Residual mutational patterns for CTE and AD are obtained by subtracting mutation profiles of age-matched controls from those of CTE and AD to show disease-specific contributions.

model; Fig. 2, E and F). We also compared the signature contributions to 27 previously reported AD neurons (*18*), in which AD showed a greater contribution of both signature A and signature C compared with age-matched controls (P = 0.037 and 0.006, respectively, LME model; Fig. 2, D and F).

To better understand the age- and disease-associated contributions of the two signatures, we calculated the ratio of each signature's total sSNVs to age-contributed sSNVs and found an elevated contribution of signature C (ratio > 1) in CTE and AD (Fig. 2G), suggesting shared mutational mechanisms in both neurodegenerative tauopathies. When we compared the relative contribution of the two signatures after adjusting for age, we found that signature C accounted for a greater proportion of the increase in CTE than in AD (Fig. 2H). META-CS data identified a mean contribution of signature C ranging from 22 to 24% in double-stranded SNVs (dsSNVs) and around 67% in single-stranded SNVs (ssSNVs) without differences between CTE and age-matched controls (fig. S6). This finding, combined with the predominant contribution of dsSNVs (~85%) in PTA neurons, suggests that a substantial amount of signature C is double-stranded. Although the precise mechanism of signature C and SBS8-like variants has not been definitively determined, the C>A mutations in this pattern are associated with alterations in reactive oxygen species (ROS) in other contexts ([29]) and represent one possible contributor to disease pathogenesis.

We also conducted de novo signature analysis using nonnegative matrix factorization (NMF) ([30]) (see methods). We identified two de novo signatures, N1 and N2, that shared many similarities with signatures C and A, respectively, but also had substantial overlaps with each other (such as SBS5 and SBS16) (fig. S7) which may limit their interpretability as they may not represent distinct biological processes. To investigate other potential contributing factors to the excess sSNVs in CTE, we aggregated mutations on the basis of their disease status (neurotypical control, CTE, and AD) and fitted them to COSMIC SBS signatures (Fig. 2I). The aging effect represented by SBS5 dominated across all three groups. However, after subtracting the pattern of age-matched controls from CTE and AD, we obtained residual patterns that were disease specific (Fig. 2I and fig. S7D), suggesting elevated contributions of SBS29 and SBS32 in CTE that are almost absent in controls. SBS29 is a signature associated with tobacco chewing, yet distinct from tobacco smoking–related SBS4 and SBS92 ([27]). A high prevalence of smokeless tobacco usage has been reported in athletes ([31], [32]); smokeless tobacco usage in American football players is unknown. SBS32 is associated with azathioprine treatment and was recently reported to be strongly associated with age in oligodendrocytes, with an absence in aging neurons ([23]). Other COSMIC signatures, including SBS30 and SBS89, were present in controls but increased in CTE; SBS30 contributed to signature C (fig. S7C) and was associated with deficient base excision repair which may also play a role in the excessive DNA damage, whereas the mechanism underlying SBS89 is unclear.

sIndels in CTE neurons

Compared with sSNVs, sIndels showed a more pronounced increase in CTE than in controls ($P = 0.004$, LME model; $P = 2.8 \times 10^{-8}$, two-tailed Wilcoxon test; Fig. 3, A and B, and fig. S2, D to F). RHI neurons showed no differences in sIndel burden versus controls, further supporting the hypothesis that increased somatic burden is specific to CTE pathology. CTE neurons accumulated a mean of 312 excess sIndels per cell compared with controls, approximately three times the CTE-specific sSNV increase and equivalent to sIndels accumulated over >100 years (given the normal neuronal rate of two to three sIndels per year per genome). These CTE excess sIndels are predominantly 2- to 4-bp (base pair) deletions (Fig. 3C). Reanalysis of PTA AD neurons from our previous study ([18]) revealed a similar phenomenon in AD (fig. S8).

We developed a new computational pipeline (see methods) specifically for identifying sIndels in single-cell META-CS data and extracted mutational signatures that represented double-stranded and single-stranded Indels (dsIndels and ssIndels) in CTE and neurotypical controls (Fig. 3D). We confirmed the robustness of these signatures with different calling thresholds (fig. S9, A to D). After fitting the dsIndel and ssIndel signatures to PTA-profiled sIndels (see methods), we found that the excess sIndels in some CTE individuals are primarily single-stranded, whereas other CTE individuals and controls have more modest numbers of sIndels that are primarily double-stranded (Fig. 3E).

To explore potential biological processes giving rise to dsIndels and ssIndels in CTE, we decomposed META-CS–derived signatures into COSMIC Indel (ID) signatures (v.3.2) (Fig. 3F). Some signatures were present in both CTE and neurotypical controls. ID5, a clocklike signature associated with aging, contributes to dsIndels and at a lesser degree to ssIndels of both groups. ID12, a signature with unknown etiology where Indels occur primarily in repetitive regions, showed a strong contribution to ssIndels of both groups, implying that repetitive regions may be prone to single-stranded lesions or sequencing errors. As the current technology may not distinguish sequencing errors or amplification artifacts from single-stranded lesions, we focused on differences observed between control and CTE that are more likely to reflect biological processes. ID4, mainly characterized by 2- to 4-bp deletions, has a predominant and robust (fig. S9, E to H) presence in CTE and is minimal in controls, consistent with the bias toward 2- to 4-bp deletions observed in CTE PTA neurons (Fig. 3C). Although the signal is more pronounced in the ssIndel signature, the dsIndel signature also showed a substantial ID4 contribution, indicating a mutagenic process where a subset of ID4-associated DNA lesions become double-stranded mutations.

Somatic ID4-like deletions in certain CTE individuals

The excess sIndels in CTE were driven by a subset of individuals (Fig. 4A), designated "High-Indel CTE," whereas other CTE individuals exhibited much lower excess sIndels compared with age-matched controls, designated "Low-Indel CTE." This distinction in excess sIndels between High-Indel and Low-Indel CTE groups remained after controlling for quality metrics (see methods and fig. S10). Although individuals in the High-Indel group were generally older (>70 years old), two younger individuals suggest that other factors beyond age may contribute. We observed modest correlation between excess sIndels and excess sSNVs in High-Indel CTE at the cellular level, but limited power precludes a clear conclusion (fig. S11). Furthermore, we found a significant association between sIndels in CTE and the duration of symptoms (the time between symptom onset and death; $P = 0.022$, LME model), which was stronger than the association with age at death ($P = 0.059$, LME model), suggesting that the duration of disease might contribute to excess sIndels. The apolipoprotein E (*APOE*) ε4 allele has been identified as a prominent risk factor for AD ([33]) and is associated with the severity of CTE tau pathology ([34]), however, we found no association of *APOE* ε4 with sIndels in CTE after controlling for age (see methods). Years of playing football—a proxy for cumulative RHI exposure—also did not show association, consistent with our findings in sSNVs. Reanalysis of PTA AD neurons ([18]) revealed similarly high rates of sIndels in certain AD individuals (designated "High-Indel AD" and "Low-Indel AD"; Fig. 4B), indicating that excessive sIndels may contribute to multiple tau-based neurodegenerative diseases.

The High-Indel CTE and AD individuals present a distinct mutational pattern. Using the counts of each context of PTA sIndels detected in each individual, unsupervised clustering identified High-Indel CTE and AD as a separate cluster from Low-Indel CTE and AD and neurotypical controls (Fig. 4C and fig. S12). The same clustering results were reinforced by META-CS data (fig. S13). The distinct pattern in High-Indel CTE and AD individuals is mainly characterized by 2- to 4-bp deletions, where the proportion of 2- to 4-bp deletions is significantly higher in High-Indel CTE than in age-matched controls for both dsIndels and ssIndels ($P = 1.0 \times 10^{-4}$ and $P = 7.5 \times 10^{-4}$, respectively, two-tailed Wilcoxon test, Fig. 4, D and E). In line with the fact that ID4 comprises primarily 2- to 4-bp deletions, we observed a consistent contribution of both double- and single-stranded ID4 across all High-Indel CTE (Fig. 4F), significantly higher than in Low-Indel CTE (fig. S14; dsIndel: $P = 0.003$, ssIndel: $P = 0.005$, two-tailed Wilcoxon test) and control groups (fig. S14; dsIndel: $P = 0.001$, ssIndel: $P = 1.0 \times 10^{-4}$, two-tailed Wilcoxon test). This highlights the role of ID4-associated biological processes in elevated DNA damage in CTE. Although the etiology of ID4 is not completely understood, recent studies ([35], [36]) demonstrated that ID4 may be mediated by topoisomerase 1 (TOP1) activity and linked to neurodegeneration.

Influence of transcriptional activity and chromatin accessibility on somatic mutation in CTE

Gene transcription and epigenetic state are closely linked to DNA damage and repair in neurons ([37], [38]). In contrast to mitotic cells,

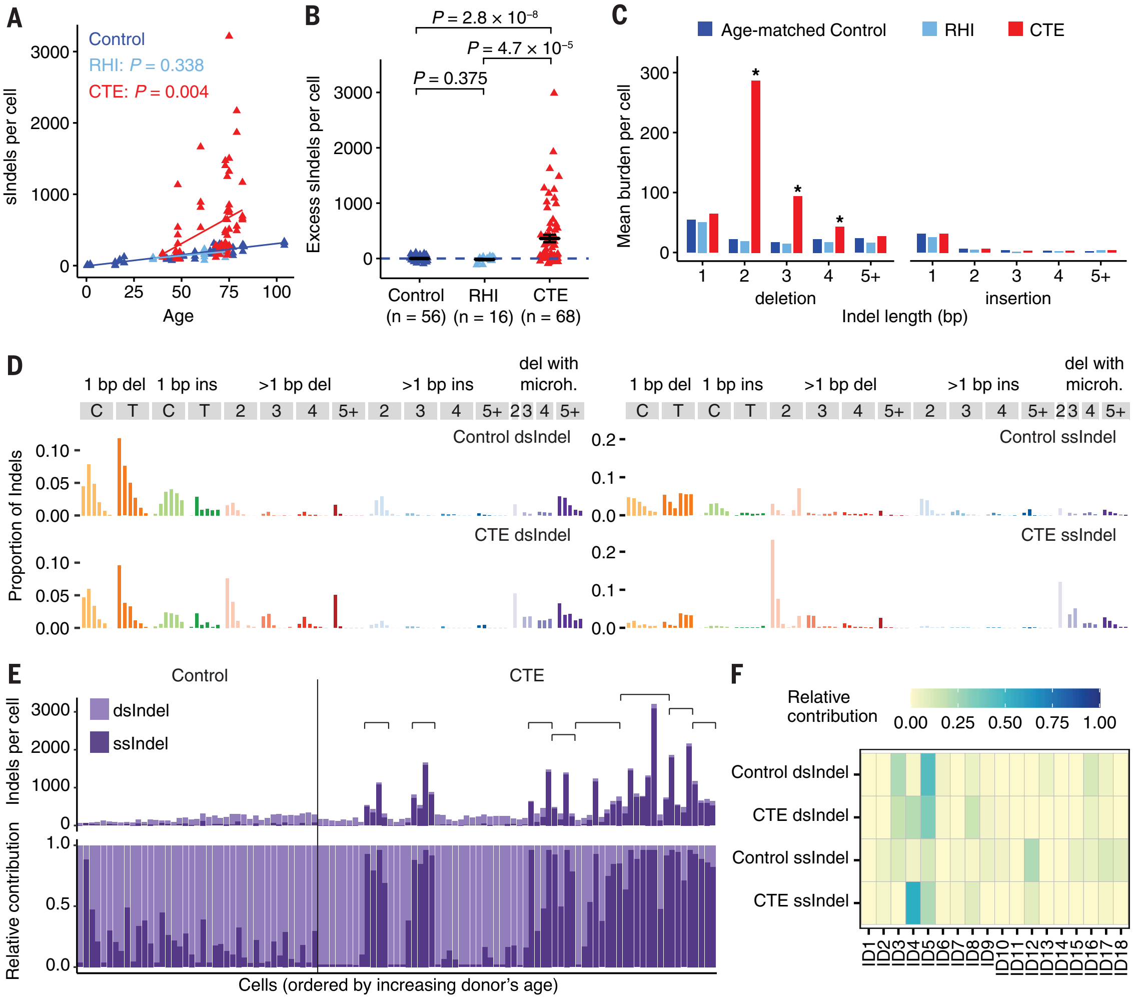


Fig. 3. sIndel burden and mutational signatures in CTE neurons. sIndel burdens were identified from PTA scWGS data, with strand-related analysis inferred from mutational signatures extracted from META-CS scWGS data. (**A**) sIndel burden in CTE (red), RHI (light blue), and neurotypical control (dark blue) brains with a significant increase in CTE but not in RHI when compared with controls. sIndel burden (from each neuron as a triangle) estimated by SCAN2 is fitted against age by clinical conditions using LME models (neurotypical control, dark blue; RHI, light blue, $P = 0.338$; CTE, red, $P = 0.004$). P values compare each clinical condition against controls. (**B**) Excess sIndel burden in RHI and CTE compared with neurotypical control after adjusting for age. Data are mean ± standard error. The dashed blue line shows sIndels attributable to age (zero excess). P values are from two-tailed Wilcoxon tests. (**C**) Comparison of all types of sIndels across age-matched controls, RHI, and CTE. Data are mean burden per cell. Asterisk denotes significant changes in certain types of sIndels when compared with age-matched controls ($P < 0.05$, two-tailed Wilcoxon test). (**D**) Double-stranded (ds) and single-stranded (ss) Indel signatures in neurotypical controls and CTE extracted from META-CS data. (**E**) Absolute (top) and relative (bottom) contribution of dsIndel and ssIndel signatures in PTA-profiled neurotypical control and CTE neurons. The pair of dsIndel and ssIndel signatures used for decomposition is determined by the clinical condition of PTA neurons. CTE cases with a pronounced ssIndel contribution are indicated by brackets. Cells with <15 Indels are not shown. (**F**) Decomposition of dsIndel and ssIndel signatures to COSMIC ID database.

replication-independent processes account for most somatic mutations in healthy neurons, where both gene expression and chromatin accessibility have been reported to be positively correlated with mutation burden ([17], [18], [23]). To investigate whether somatic mutations have distinct enrichment patterns in CTE neurons, we generated single-nucleus RNA sequencing (snRNA-seq) data from CTE and neurotypical control samples (fig. S15) and used published single-nucleus assay for transposase-accessible chromatin with sequencing (snATAC-seq) data from sample-matched neurotypical controls ([23]). We extracted neuronal profiles to eliminate potential biases from other cell types.

We found somatic mutations significantly more enriched in transcribed genes than in nontranscribed genes in control and CTE ($P <$ 0.05 for all, two-tailed Wilcoxon test), which aligns with previous findings and highlights the role of transcriptional activity in neuronal mutagenesis. We further confirmed positive correlations of somatic mutations with gene expression and chromatin accessibility in both control and CTE (Fig. 5A). Using signatures to further dissect enrichment patterns according to mutational processes, we observed positive correlations between gene expression and sSNVs attributed to signature A for both control and CTE, whereas signature C showed a negative correlation (Fig. 5B), consistent with previous results in AD neurons ([18]) and reinforcing the shared findings between CTE and AD. Chromatin accessibility showed no clear trend for signature C–related sSNVs, which may be explained by complex relationships between chromatin accessibility and gene expression activity in highly expressed genes (fig. S16) as well as potential confounds of global genome NER (GG-NER)–related sSNVs in regulatory elements, as signature C was identified in genetic disorders with deficiency in either TC-NER or GG-NER ([21]). For sIndels, we separately extracted double- and single-stranded signatures from High-Indel CTE, Low-Indel CTE, and controls and found that total burden as well as burden attributed to either signature showed moderate positive correlations with both

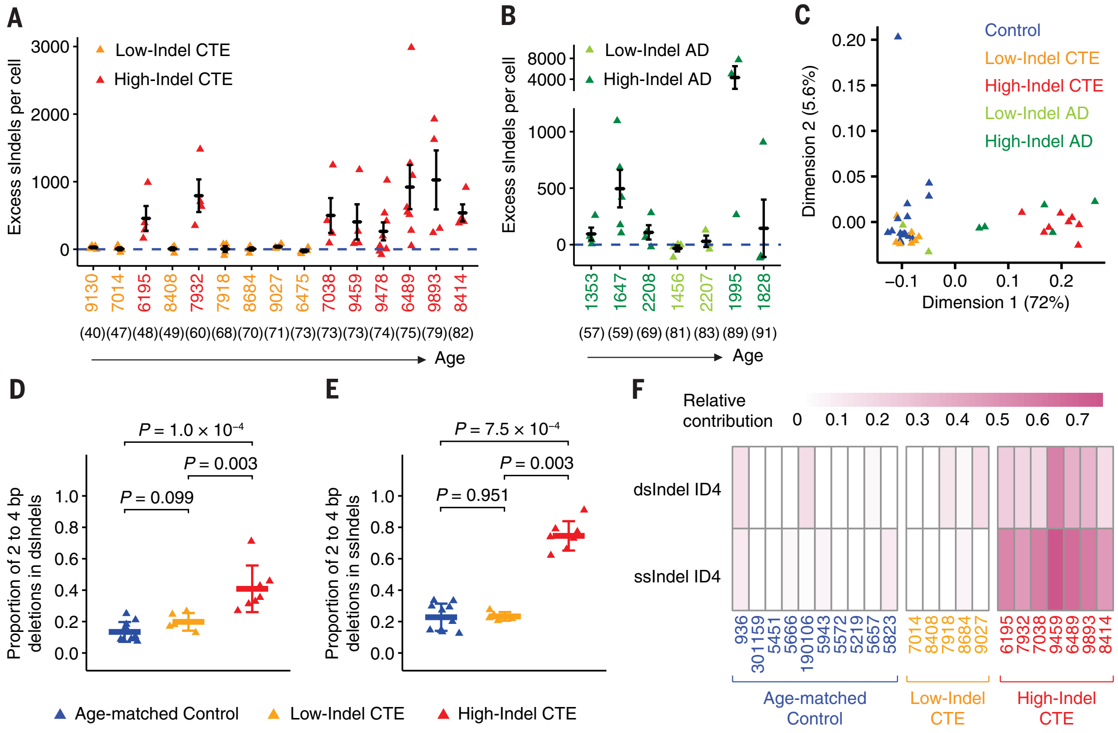


Fig. 4. Somatic ID4-like deletions in certain individuals with CTE. (**A** and **B**) Excess sIndels in each CTE (A) and AD (B) case ordered by increasing age (indicated in parentheses). CTE and AD cases are separated into either the High-Indel group (High-Indel CTE, red; High-Indel AD, dark green) or the Low-Indel group (Low-Indel CTE, yellow; Low-Indel AD, light green) on the basis of whether they have an excess of sIndels (>50). Data are mean ± standard error. The dashed blue line shows sIndels attributable to age (zero excess). (**C**) Principal components analysis (PCA) clustering of sIndels from each case across neurotypical controls, CTE, and AD. sIndels from each case were aggregated and stratified into 83 contexts defined by COSMIC. The first two dimensions are used for visualization with the percentage of variance explained shown for each dimension. Cases with <15 sIndels are not shown. (**D** and **E**) Comparison of proportions of 2- to 4-bp deletions in dsIndels (D) and ssIndels (E) across age-matched controls, Low-Indel CTE, and High-Indel CTE. Each triangle represents an individual from META-CS data. High-Indel CTE versus age-matched controls: $P = 1.0 \times 10^{-4}$ and $P = 7.5 \times 10^{-4}$, two-tailed Wilcoxon test. Data are mean ± standard deviation. (**F**) Relative contribution of ID4 to dsIndels and ssIndels of each case from META-CS data shown as a heatmap. Case IDs are colored by their group assignment (age-matched control, dark blue; Low-Indel CTE, yellow; High-Indel CTE, red).

gene expression and chromatin accessibility (Fig. 5, C and D). The similar pattern between dsIndel and ssIndel suggests that they share similar mutagenic mechanisms, with single-stranded lesions representing an intermediate stage before double-stranded mutations.

We further examined the distribution of somatic mutations in protein-coding genes for their potential functional impact. Gene Ontology (GO) analysis of genes mutated by sSNVs (Fig. 5E) and sIndels (Fig. 5F) revealed consistent enrichments in neuronal functions such as synaptic structures and neuron projection. This observation aligns with our finding that somatic mutations are enriched in regions of high expression and open chromatin, reflecting the active transcription of neuronal genes in neurons. Taken together with the CTE-specific somatic burden increase, our results suggest that excess somatic mutations, notably sIndels in CTE neurons, may frequently disrupt essential neuronal genes and eventually contribute to neurodegeneration (fig. S17).

Discussion

In this study, we characterized somatic mutations in CTE neurons using two different scWGS methods and revealed elevated DNA damage leading to sSNVs and sIndels in CTE, in patterns distinct from normal aging. We propose specific mutagenic pathways for the accumulation of sSNVs and sIndels (fig. S17). Excess sSNVs in CTE were largely contributed by signature C, a pattern that is rare in controls and elevated in AD. Given the shared feature of tau pathology, CTE and AD may share common pathways for accumulating DNA damage, possibly related to oxidative stress (18). We also found a subset of High-Indel CTE individuals harboring excess double- and single-stranded Indels, with a predominant contribution from ID4 (27). Analysis of AD individuals showed a subset with similar excessive sIndels, which was also seen in a parallel study examining tau pathologic states (39).

Previous studies have highlighted RHI as the likely cause of CTE (40, 41); however, not all individuals exposed to RHI develop CTE. We found that the somatic mutation burden of the RHI group was similar to that of neurotypical controls, implying that the elevated DNA damage in CTE is distinct from long-term exposure to RHI. A limitation of this study is that the RHI group is small with a narrow age range, and a larger group of varying ages and RHI exposures is needed to validate this finding. Our cohort focused on severe CTE (stages III and IV); future studies that include early-stage CTE could assess the role of somatic mutations in disease progression. Moreover, other distinctive features of CTE, such as the perivascular localization of tau deposition at the depths of the cortical sulci, may reveal previously unknown associations between DNA damage and pathogenesis. Our analysis between the sulcus and gyrus showed no difference (fig. S4), but statistical power was limited. In addition, we did not detect any somatic mutations in genes known to be associated with neurodegenerative tauopathies (*APP*, *PSEN1*, *PSEN2*, *APOE*, and *MAPT*), similar to a previous report in AD (18).

Duplex sequencing data distinguished single-stranded and double-stranded genomic alterations, showing that sSNVs and sIndels in

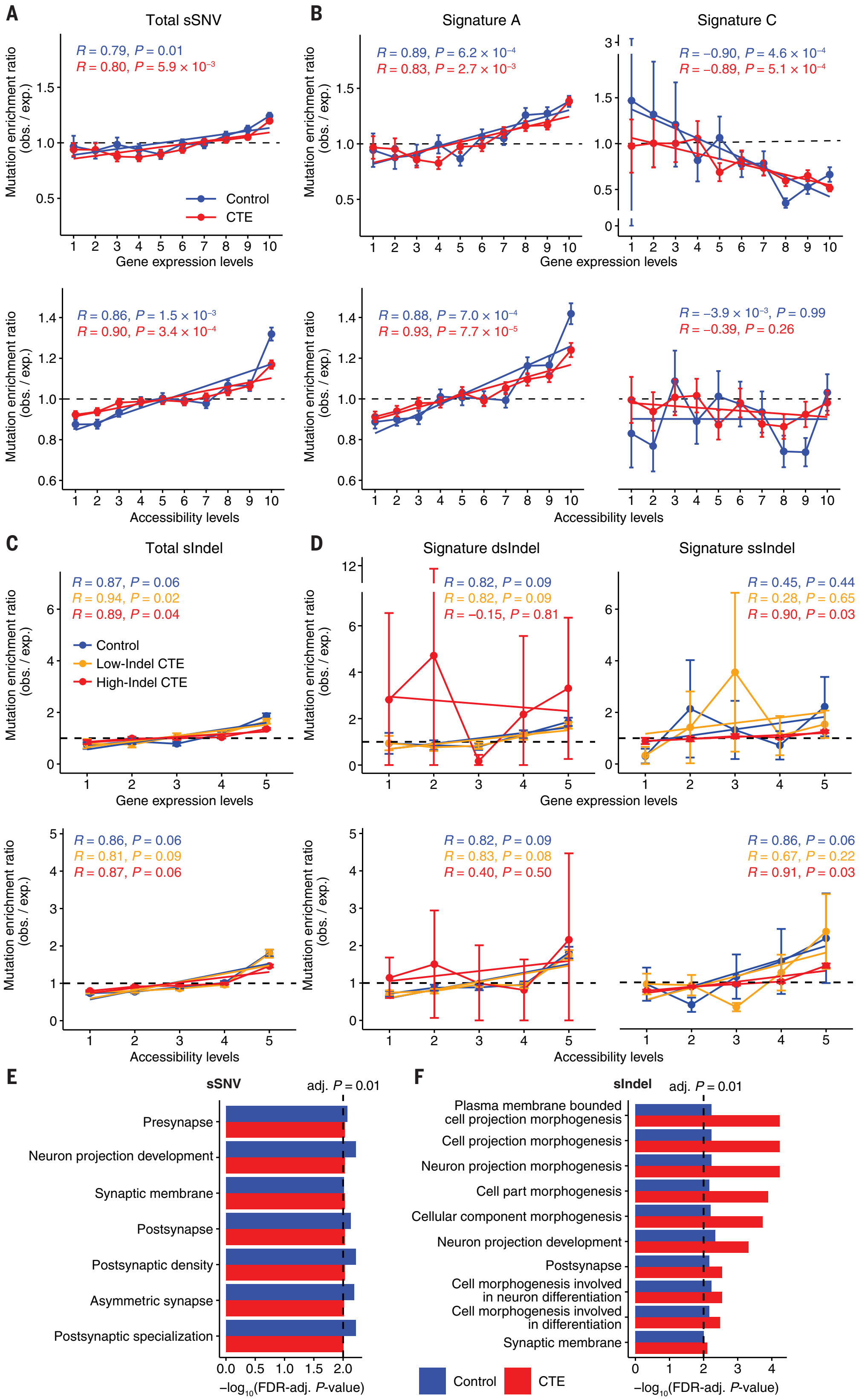


Fig. 5. Enrichment analysis of sSNVs and sIndels in CTE neurons. (**A** to **D**) sSNV [(A) and (B)] and sIndel [(C) and (D)] density as a function of gene expression and chromatin accessibility levels. Neuronal transcriptional profiles were characterized from snRNA-seq data sequenced in this study. Neuronal accessibility profiles were obtained from snATAC-seq data in Ganz *et al*. ([23]). Genes and open chromatin regions are separated into 10 (for sSNV) or 5 (for sIndel) equally sized groups with increasing levels indicating increasing expression or accessibility. Observed density of each expression or accessibility group is obtained by overlapping the original mutation call sets with regions of each group. Expected density is calculated from 1000 permutations of mutation call sets overlapped with regions of each group. Enrichment ratio is calculated by

observed/expected density for each permutation, and mean ratio over 1000 permutations is used to construct the trend line by linear model (error bar indicates standard deviation). Pearson correlation coefficient (*R*) and two-tailed *P* value (*P*) are shown. obs., observed; exp., expected. (**E** and **F**) Gene ontology (GO) analysis of genes where sSNVs (E) and sIndels (F) are located. GO terms with false discovery rate (FDR)–adjusted $P < 0.01$ in both CTE and control are reported.

CTE neurons appeared to have distinct distributions of double- and single-stranded events that may be specific to each individual. Most sSNVs from all CTE individuals were represented on both strands; however, High-Indel CTE individuals present a specific pattern contributed by many features of ID4. ID4 was recently reported to occur when ribonucleotides are misincorporated into DNA and subsequently cleaved by TOP1 instead of RNase H2 (*35*). These double-stranded deletions are thought to arise via an intermediate step of single-stranded >1-bp deletions, which aligns with our observation that ID4-like patterns have a stronger signal in ssIndels. Although the association between RNase H2 deficiency and neurodegenerative disease is unknown, our results raise the possibility that a large quantity of genome-embedded ribonucleotides in CTE and AD neurons might overwhelm the ribonucleotide excision repair pathway, with the TOP1-mediated process leaving large numbers of >1-bp deletions on single DNA strands of which only a small proportion could be efficiently repaired or fixed into double-stranded Indels in postmitotic neurons (*36*). A parallel study (*36*) found a similar sIndel pattern in two other neurodegenerative diseases with distinct pathologies and provided direct experimental evidence of some expected features of TOP1-mediated mutagenesis, suggesting an even wider impact of TOP1 activity in neurodegeneration. It is also worth noting that ID4 may not fully explain the DNA damage in CTE, as some excess sIndels, particularly 2-bp deletions in one repeat unit, are limited in ID4. This suggests that other mutagenic processes may play a role in congruence with or independent of ID4-associated processes. Nonetheless, our analysis highlights that CTE can be viewed as a genomic process in affected neurons, with progressive transcription-related accumulation of sSNVs and especially sIndels—which accumulate to numbers equivalent to >100 years of excess aging—creating the potential for severe dysregulation of the neuronal transcriptome (fig. S17).

Materials and methods

Data reporting

No statistical methods were used to predetermine sample size. Experimenters were not blinded, and experiments were not randomized.

Human tissue samples and CTE case selection

Postmortem frozen human brain tissues were obtained from the UNITE Brain Bank (formerly VA-BU-CLF) at Boston University, the Massachusetts Alzheimer's Disease Research Center (MADRC) at Massachusetts General Hospital, and the NIH Neurobiobank at the University of Maryland Brain and Tissue Bank (UMBTB).

Tissue samples were collected and distributed for research and publication under the protocols approved by the Boston University–Veterans Affairs Institutional Review Board (for BU-VA: S07-02-0087), MassGeneral Brigham Institutional Review Boards (for MADRC: 1999P009556/MGH, expedited waiver category 5), and the University of Maryland Institutional Review Board (for UMBTB: 00042077) and after provision of written authorization and informed consent. Research on these de-identified specimens and data was performed at Boston Children's Hospital with approval from the Committee on Clinical Investigation (S07-02-0087 with waiver of authorization, exempt category 4) and at Brigham and Women's Hospital with approval from the MassGeneral Brigham Institutional Review Board (2019P003790 for secondary use as nonhuman subjects research).

Neurotypical control and AD cases were included from part of previous studies (*17*, *18*, *23*). Neurotypical control cases had no clinical history of neurological disorders, and AD cases were pathologically confirmed Braak stage V-VI. CTE cases were pathologically confirmed stage III-IV (*3*, *42*). Neither AD nor CTE cases had other notable neurodegenerative pathology.

Isolation of individual pyramidal neurons using FANS

Isolation of single neuronal nuclei was performed through fluorescence-activated nuclear sorting (FANS) for the neuronal nuclear transcription factor (NeuN), as described in previous work (*18*, *21*, *43*). In brief, nuclei were prepared from unfixed frozen (at −80°C) human brain tissue in a dounce homogenizer using a chilled tissue lysis buffer [10 mM Tris-HCl, 0.32 M sucrose, 3 mM $Mg(OAc)_2$, 5 mM $CaCl_2$, 0.1 mM EDTA, 1 mM dithiothreitol (DTT), 0.1% Triton X-100, pH 8] on ice. Lysed tissue was then layered on top of a sucrose cushion buffer [1.8 M sucrose 3 mM $Mg(OAc)_2$, 10 mM Tris-HCl, 1 mM DTT, pH 8] and ultracentrifuged for 1 hour at 30,000*g*. Nuclear pellets were resuspended in ice-cold phosphate-buffered saline supplemented with 3 mM $MgCl_2$, filtered, then stained with DAPI and anti-NeuN (RBFOX3) antibody directly conjugated to Alexa Fluor 488 (AF488) (Millipore MAB377X, clone A60, 1:1,250; RRID: AB_2149209). DAPI staining allowed for isolation of single diploid neuronal nuclei apart from debris and doublet droplets. Previous work showed that using flow cytometry (with software BD FACSDiva v.8.0.2) to gate for the nuclei population with the highest NeuN signal as well as the highest forward scatter area (FSC-A) signal produced 99.3% excitatory pyramidal neurons (*18*).

scWGS of pyramidal neurons using PTA

Isolated single nuclei were sorted one nucleus per well into 96-well plates and their genomes were amplified by PTA (*22*). The quasi-linear process of amplification in PTA has been found to reduce amplification artifacts and improve variant calling (*17*). PTA was performed using the ResolveDNA Whole Genome Amplification Kit (BioSkryb Genomics) as described in previous work (*18*). In brief, nuclei were sorted into 3 μl Cell Buffer prechilled on ice and lysed by addition of 3 μl MS Mix, with mixing at 1400 rpm performed after each step. Lysed nuclei were then neutralized with 3 μl SN1 buffer, followed by 3 μl of SDX reagent, a 10-min incubation at room temperature, and 8 μl of reaction mix (containing polymerase) for a 20 μl of total reaction volume. Next, amplification was performed at 30°C for 10 hours, followed by enzyme inactivation at 65°C for 3 min and DNA cleanup using AMPure. The yield was then determined by PicoGreen binding (Quant-iT dsDNA Assay Kit, Thermo Fisher Scientific). Multiplex polymerase chain reaction (PCR) at four random genomic loci was used as quality control (*21*), where samples with positive amplification at all four loci were retained for Illumina sequencing.

Library preparation followed a modified KAPA HyperPlus Library Preparation protocol described in the ResolveDNA EA Whole Genome Amplification protocol. In brief, end repair and A-tailing were performed for 100 to 500 ng amplified DNA input, followed by adapter ligation using the SeqCap Adapter Kit (Roche, 07141548001) and cleanup using AMPure. After on-bead PCR amplification, size selection was carried out using AMPure for libraries with a size of 300 to 600 bp. Quality control was performed on selected libraries using PicoGreen and TapeStation HS DS100 Screen Tape (Agilent PN 5067-5584), followed by paired-end sequencing (150 bp × 2) on Illumina NovaSeq sequencers at 30× coverage (table S2). Data from PTA-amplified neuronal genomes in CTE and RHI were analyzed alongside previously reported data from control (*17*, *23*) and AD neurons (*18*).

scWGS of pyramidal neurons using modified META-CS

The genomes of single neuronal nuclei were amplified by a modified version of META-CS (*19*), a transposase-based whole genome amplification technique in which each DNA fragment is tagged and barcoded with 16 unique tags [dataset S1 of Xing *et al.* (19)], allowing for single-cell, strand-resolved identification.

DNA oligos were ordered from IDT. Each of the 16 META-CS transposons were annealed and assembled into transposomes with Diagenode Tagmentase (Diagenode; C01070010) per manufacturer's protocol and stored at −80°C.

Single neuronal nuclei, isolated as previously described, were sorted one per well into 96-well plates and lysed in 2 μl of 1X Single Cell Lysis Buffer [20 mM Tris, pH 8.0, 20 mM NaCl, 0.15% Triton X-100, 25 mM DTT, 1 mM EDTA, 1.5 mg/ml Thermolabile Proteinase K (TLPK) (NEB, P8111S)] at 30°C for 1 hour, 55°C for 10 min. Single-cell lysates were stored at −20°C if not immediately amplified.

Lysed nuclei were then transposed by the addition of 8 μl transposition mix (5 μl Diagenode 2X Tagmentation buffer (Diagenode; C01019043), 1 μl diluted META-CS transposome, 2 μl H_2O), mixed at 1640 rpm for 1 min, spun down at 1500 rpm, and incubated at 55°C for 15 min. Transposases were removed by the addition of 2 μl 6X stop buffer containing 300 mM NaCl, 45 mM EDTA, 0.01% Triton X-100, and 1 mg/ml TLPK, with mixing and incubation at 37°C for 30 min, 55°C for 10 min.

First-strand tagging was performed by the addition of 13 μl Strand Tagging Mix 1 containing 5 μl Q5 reaction buffer (NEB; B9027S), 5 μl Q5 high GC enhancer (NEB; B9028A), 0.85 μl 100 μM (total) Adp1 primer mix [dataset S1 of Xing *et al.* (19)], 0.6 μl 100 mM $MgCl_2$, 0.55 μl water, 0.5 μl 10 mM (each) dNTP mix (Thermo Scientific; R0192), 0.25 μl of 20 mg/ml bovine serum albumin (NEB; B9000S), and 0.25 μl Q5 DNA polymerase (NEB; M0491S), followed with mixing and incubation 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 with 1 μl Thermolabile Exonuclease I (NEB; M0568L), with mixing and incubation at 37°C for 15 min, 65°C for 5 min.

Second-strand tagging was performed by the addition of 4 μl Strand Tagging Mix 2 containing 1 μl Q5 reaction buffer, 1 μl Q5 high GC enhancer, 0.95 μl 100 μM (total) Adp2 primer mix [dataset S1 of Xing *et al.* (19)], 0.85 μl water, 0.1 μl 10 mM (each) dNTP mix, and 0.1 μl Q5 DNA polymerase, and incubation at 72°C for 3 min, 98°C for 30 s, 62°C for 5 min, 72°C for 1 min. Adp2 primers were removed by repeating the exonuclease step described above.

Strand tagging products were amplified by the addition of 19 μl PCR mix containing 1 μl NEBNext Multiplex Oligos Universal Primer, 1 μl NEB Index Primers (NEB; E7335S, E7500S, E7710S, E7730S), 4 μl Q5 reaction buffer, 4 μl Q5 high GC enhancer, 0.4 μl 10 mM (each) dNTP mix, 8.4 μl water, and 0.2 μl Q5 DNA polymerase and incubation at 98°C for 20 s, 13 cycles of [98°C for 10 s, 72°C for 2 min], 72°C for 2 min.

Single-cell libraries were pooled together, and then libraries were purified by DNA Clean and Concentrator-5 columns (Zymo; D4013) and amplification efficiency was checked for fragment size and concentration by Agilent TapeStation or Agilent Bioanalyzer. Size selection was performed with Ampure XP beads (Beckman Coulter; A63880), wherein the pooled library was divided into three groups on the basis of fragment size. Medium-size fragments (~300 to 1000 bp) were selected first by the addition of 0.6X beads then by a further addition of 0.25X beads (for a final 0.75X).

snRNA-seq of CTE and control samples

Single-nucleus RNA sequencing (snRNA-seq) was performed on representative tissue samples (control individual UMB1465, prefrontal cortex; CTE individual CTE9130, prefrontal cortex) to assess mutational enrichment with gene expression. Isolation of nuclei was performed as described above with the following modifications: both 0.2 U μl^{-1} Protector RNase inhibitor (Roche RNAINH-RO) and 0.2 U μl^{-1} SUPERase-In RNase inhibitor (Invitrogen) were added to the tissue lysis buffer and to the immunostaining buffer, and $MgCl_2$ was omitted from the immunostaining buffer. For each sample, ~16,000 nuclei were sorted into one well of 96-well plates. snRNA-seq was performed using the 10X Genomics Next GEM Single Cell 3′ GEM Kit v3.1 and Chromium Controller, followed by Illumina sequencing.

Read mapping and BAM file generation for bulk and PTA data

BWA (v.0.7.15) (44) was first used to map reads from bulk WGS and PTA scWGS data onto the human reference genome (GRCh37 with decoy) with default parameters. Then, duplicate reads were marked by MarkDuplicates of Picard (v.2.8.0), followed by local Indel realignment and base quality score recalibration using Genome Analysis Toolkit (GATK) (v.3.5) (45) to generate BAM files for mutation calling.

Quality measures of single-cell genome amplification

To evaluate the quality of single-cell genome amplification for both PTA and META-CS data, we used a number of measures. First, we calculated the median absolute pairwise differences (MAPD) to quantify the evenness of amplification as described previously (46). MAPD score was computed by binning the genome, estimating copy number of each bin, and taking the median of absolute pairwise differences between log2-transformed copy number ratios of adjacent bins. A lower MAPD score indicates more-even amplification of the genome. Further, to account for the variance of the copy number ratio distribution, we calculated the coefficient of variation (CoV) of these absolute pairwise differences by taking the ratio of their standard deviation to their mean. Sequencing depth was estimated using the total number of properly mapped and paired reads (from samtools stats) multiplied by read length and divided by the whole genome length. To account for amplification bias, we estimated the allelic and locus dropout rates using a set of high-quality germline heterozygous SNPs that overlap with common variants from the 1000 Genome Project. Allelic dropout sites have either reference (REF) allele depth or nonreference (ALT) allele depth < 2. Locus dropout sites have total depth < 5. Strand dropout rate was estimated as the square root of allelic dropout rate.

Calling of sSNVs and sIndels from PTA data

We used Single Cell ANalysis 2 (SCAN2, v.1.0) (17) to identify sSNVs and sIndels from single-cell PTA data with matched bulk data. First, we generated four cross-sample panels. One panel for 17 control individuals from UMBTB, one panel for 4 RHI individuals, one panel for 15 CTE individuals and 2 control individuals from UNITE Brain Bank (formerly VA-BU-CLF), and one panel for 7 AD individuals. Each panel was configurated by running "scan2 config–analysis makepanel" with the following reference parameters: human reference genome GRCh37 with decoy (--ref), dbSNP v138 common variants (--dbsnp), and 1000 Genomes Phase 3 SHAPEIT2 phasing panel (--shapeit-refpanel). Each panel was then built by running "scan2 makepanel." After panel generation, mutation calling was performed for each individual by running "scan2 config–analysis call_mutations" with the same reference parameters above, each PTA BAM (--sc-bam), matched bulk BAM (--bulk-bam), and the corresponding cross-sample panel (--cross-sample-panel), followed by "scan2 run." For each single cell, SCAN2 generated both mutation calls and genome-wide somatic mutation burden estimations (autosomes only) which were used for the subsequent PTA analyses described below. Two AD cells were excluded from this study either because of failed SCAN2 run (ALZ1647BA9-C) or large discrepancy in burden estimations between SCAN2 and LiRA (1995P_201001E3), the tool originally used to profile these cells (18).

Preprocessing and filtering of META-CS data

Our pipeline to preprocess META-CS data and perform sSNV calling shares a core workflow with the previously reported method (19) and includes extra steps to further strengthen accuracy and remove false positives, which better accommodates our modified experimental protocol. First, paired-end reads were preprocessed by pre-meta including identifying transposon barcodes, merging overlapping read ends, and trimming Illumina adapters. Then, reads were mapped to the human reference genome (GRCh37 with decoy) to generate BAM files using BWA-MEM (v.0.7.17) (47) and Minimap2 (v.2.12) (48) for SNV and BWA-MEM (v.0.7.17) (47) for Indel. Of note, because each original DNA fragment was tagged by a pair of transposon barcodes, we split BAM files by barcode pairs before running mutation calling in each barcode pair BAM. This allows us to filter out reads without matching barcodes and ensures that calling was performed on a single-molecule level (for each allele). In addition, given the relatively small number of unique barcodes, there is a chance of barcode collision where different DNA fragments are tagged by the same barcode pair. Therefore, we extracted Tn5 cut sites from each read pair, with the assumption that reads amplified from the same DNA fragment should share the same Tn5 cut sites. Other quality metrics were obtained in the same way as PTA data (described above). Owing to batch effects and imbalanced coverage among pooled cells, cells with an average insert size of <280 or >500, or a standard deviation of insert size >750, were filtered out. In addition, some cells had abnormally high numbers of calls, and we found a significant overlap (~93%) between their prefiltered call sets and the gnomAD (49) SNPs of ≥1% population frequency. These cells were removed owing to potential contamination with another individual.

Calling of sSNVs from META-CS data

We generated sSNV candidates by identifying variants that have no ALT allele read from bulk but at least four total ALT reads and at least two ALT reads from each strand for duplex support in the cell. To achieve robust and accurate calling, we filtered out candidate sites that satisfied any of the following criteria: overlapping with the low-quality regions as previously reported (19), overlapping with gnomAD (49) SNPs of ≥1% population frequency, within 100 bp from another candidate site. Passing variants need to have at least four ALT reads in total with at least two ALT reads from each strand (a4s2) as well as a VAF = 1 at single-molecule level.

Calling of sIndels from META-CS data

We established a new pipeline (https://github.com/gldong/duplex-indel) for sIndel calling from META-CS data by introducing novel modules to expand on the sSNV pipeline (manuscript in preparation). First, as the majority of false positive calls originated from incorrectly merged read pairs during preprocessing, we created a module to tag the merged reads and calculate a genomic window where the merging occurred. Second, we generated an additional BAM file without read merging and filtered out any candidate sites that either overlapped with the read merging window or were not present in the BAM without read merging. In addition, we removed any candidate sites that were adjacent to a germline Indel (located within the flanking region of the germline Indel start site; flanking region is defined by 5 bp or twice of the germline Indel length on both sides, whichever is larger). Other filtering criteria were the same as the ones used for sSNV calling, except that we used gnomAD (49) Indels of ≥1% population frequency. Passing variants need to have at least four ALT reads in total with at least two ALT reads from each strand (a4s2) as well as a VAF = 1 at single-molecule level.

Calling of single-stranded variants from META-CS data

Single-stranded SNV and Indel calls were detected in a similar fashion as double-stranded calls described above. We generated candidate sites by identifying variants that have no ALT read from bulk, at least four ALT reads from the variant strand in the cell, and no ALT read from the non-variant strand in the cell. Then, the same filtering strategy except for the ALT allele balance was implemented to generate the final call set. Passing variants need to have at least four ALT reads from the variant strand and at least four REF reads from the nonvariant strand.

Additional filtering of SNV and Indel calls from META-CS

To account for batch effects and technical artifacts, we implemented a set of filters to remove false positives. The same filters were applied to both double- and single-stranded SNVs and Indels. We removed variant sites that meet at least one of the following criteria: (i) variant site is covered by more than one barcode-pair, (ii) barcode-pair covering the variant site has more than one set of Tn5 cut sites, (iii) variant site located within 20 bp of read ends.

Statistical models of somatic mutation burden

We used linear mixed-effects (LME) models from the lme4 (v.1.1-27.1) R package (50) to investigate potential associations between somatic mutation burden and other covariates of interest. Somatic mutation burden was modeled as a continuous outcome, covariates of interest including age, clinical status, and cell quality measures were modeled as fixed effects, and individuals were modeled as random effects due to potential correlations between cells from the same individual. *P*-values of fixed effects were obtained from *t* tests with the Satterthwaite approximation using the lmerTest (v.3.1-3) R package (51). Each clinical group was tested against controls separately to obtain the corresponding *P* value. To calculate QC-corrected burden, we first modeled each quality measure and age as fixed effects in control neurons to obtain the coefficient of this quality measure's contribution and then subtracted its contribution from total somatic burden of each cell before modeling the corrected somatic burden against age and clinical status to generate QC-corrected burden comparisons.

As disease cells often exhibit large variation in somatic mutation burden (which could also represent biological effects), the various assumptions of the LME model may not always hold. Therefore, to strengthen our statistical testing results, we performed additional nonparametric Wilcoxon tests on excess mutation burdens across clinical groups after adjusting for age. To calculate the age effects on mutation burden, we fitted burdens from control cells in an LME model with age as fixed effects and individuals as random effects, and checked the model assumptions (linearity, homoscedasticity, and normality) (fig. S2). Then we used the coefficient of age to calculate an expected age-contributed mutation burden in cells across clinical groups, before subtracting it from the cell's total mutation burden to obtain excess mutation burden. Finally, we used two-tailed Wilcoxon test, a nonparametric test that does not require those assumptions for LME, to test the difference in excess mutation burdens between pairs of clinical groups.

Association analysis of somatic mutation burden

We used the same LME model described above to evaluate the association between somatic mutation burden and covariates of interest in CTE individuals. For years of playing football, sSNV or sIndel burden was modeled as a continuous outcome, age and the number of years playing football were modeled as fixed effects, and individuals were modeled as random effects. *P* values of fixed effects were obtained from *t* tests with the Satterthwaite approximation using the lmerTest (v.3.1-3) R package (51). For *APOE* ε4, we first genotyped all CTE individuals at two SNPs, rs429358 and rs7412, using their bulk WGS data (six possible genotypes are ε2/ε2, ε2/ε3, ε2/ε4, ε3/ε3, ε3/ε4, and ε4/ε4). Then, we used three genetic models of ε4 allele, dominant, additive (with linear penetrance), and recessive, to test the association between *APOE* ε4 genotype and somatic burden. Somatic burden was modeled as a continuous outcome, age and *APOE* ε4 genotype were modeled as fixed effects, and individuals were modeled as random effects. *P* values of fixed effects were obtained from *t* tests with the Satterthwaite approximation using the lmerTest (v.3.1-3) R package (51).

Double- and single-stranded signature analysis

Strand-specific signatures were extracted from META-CS call sets and normalized. To better interpret these signatures, we decomposed them to the COSMIC v3.2 database (https://cancer.sanger.ac.uk/signatures) with matched mutation type using MutationalPatterns (v.3.0.1) R package (30).

Because these signatures were derived independently, the double- and single-stranded signatures for each mutation type exhibited collinearity that made them difficult for refitting. Therefore, two intermediate signatures were introduced to serve as a proxy for refitting. For each mutation type, the double- and single-stranded signatures, DS and SS, were combined and normalized, from which two intermediate signatures, IM_1 and IM_2, were extracted using NMF. DS and SS were then refitted to IM_1 and IM_2 so that they could be reconstructed as

$$\widehat{DS} = a_1 \times IM_1 + a_2 \times IM_2,$$
$$\widehat{SS} = b_1 \times IM_1 + b_2 \times IM_2$$

where a_1, a_2, b_1, and b_2 were proportional contributions normalized to IM_1 and IM_2 (i.e., $a_1 + b_1 = 1$ and $a_2 + b_2 = 1$).

To estimate the proportions of double- and single-stranded calls within the PTA call sets, for each mutation type in each cell, PTA calls were fitted to the intermediate signatures

$$\widehat{PTA} = c_1 \times IM_1 + c_2 \times IM_2$$

and the signature contributions were converted to describe DS and SS

$$\widehat{PTA}_{\text{converted}} = k_1 \times DS + k_2 \times SS$$

where $k_1 = c_1 \times a_1 + c_2 \times a_2$ and $k_2 = c_1 \times b_1 + c_2 \times b_2$. Note that all signatures and spectra are vectors.

Known and de novo mutational signature analysis

To study the contributions of known and de novo mutational signatures to the PTA call sets, we first categorized the mutation calls from each cell into context groups predefined by COSMIC v3.2 (96 contexts for SNVs and 83 contexts for Indels). For sSNVs, we used MutationalPatterns (30) to fit the PTA calls to known signatures A and C reported in Lodato *et al.* (21) to get the proportional contribution of each signature to each cell. Then, signature-specific somatic mutation burden was calculated by multiplying the contribution by the genome-wide burden, which is comparable between cells. De novo signatures were extracted using the NMF-based method in MutationalPatterns. The number of signatures to extract was determined with 200 NMF runs based on the Kullback-Leibler divergence (method = "brunet"). PTA calls were fitted to the de novo signatures in the same way as for known signatures. We used the COSMIC v3.2 database to decompose both known and de novo signatures for further interpretation.

Permutation of sSNV and sIndel calls

Permutation sets are crucial for robust enrichment analyses by controlling for potential biases that exist in the original mutation calls and providing powered statistical tests to evaluate the significance of enrichment. For each permutation, the original calls were randomly shuffled across the callable regions of each cell while preserving their chromosome and trinucleotide context. We generated 1000 permutation sets for each mutation type in each cell using SCAN2 by running "scan2 config–analysis permtool" with original mutation calls (--permtool-muts), the human reference genome GRCh37 with decoy (--permtool-bedtools-genome-file), and 1000 permutations (--permtool-n-permutations), followed by "scan2 permtool."

Droplet-based snRNA-seq analysis

Gene count matrices for both control and CTE were acquired by aligning reads to the GRCh37 genome (v.3.0.0) using Cellranger (v.6.1.0) (52) with default parameters. Then we used the standard workflow from Seurat (v.4.0.5) (53) to process the snRNA-seq data. For the purpose of quality control, we removed genes that were not expressed in <3 cells and cells with (i) <200 genes, (ii) <500 counts or >15,000 counts, or (iii) mitochondrial gene percentage of >5%. After applying these filters, we obtained 4619 cells for the control dataset and 4080 cells for the CTE dataset. Next, the data were normalized using the function "LogNormalize" with "scale.factor" = 10000 and scaled to a maximum value of 10. We then performed dimension reduction (principal components analysis, t-distributed stochastic neighbor embedding, and uniform manifold approximation and projection) and cell clustering using Louvain (54). To identify the cell types that were captured in the snRNA-seq data, we selected marker genes for each cluster by statistical tests and assigned cell-type labels according to previously reported marker lists (55). Through visual evaluation, cells with the same identity were clustered together.

Enrichment analysis with snRNA-seq and snATAC-seq data

We performed enrichment analysis of somatic mutations in transcribed gene or open chromatin regions using our in-house snRNA-seq data and previously reported snATAC-seq data generated from matching control samples (23). We extracted the gene expression and chromatin accessibility profiles of neurons annotated in snRNA-seq and snATAC-seq data to match the PTA data.

For gene expression enrichment analysis, we used the expression profile of control for mutation calls from control cells, and the expression profile of CTE for mutation calls from CTE cells. First, genes were ranked on the basis of their expression and divided into equal-sized groups (10 groups for sSNVs, 5 groups for sIndels). Then we intersected both original and permuted mutation calls with gene regions of each expression group. Gene regions were annotated by ANNOVAR (56) using the database GRCh37 refGene (57). Mutation calls that overlapped with multiple genes were removed. A ratio of observed to expected number of calls (i.e., enrichment ratio) was calculated for each permutation round where the observed number came from the original call set and the expected number came from the permuted call set. Enrichment was reported as the mean and standard deviation of the 1000 ratios. To test enrichment differences between transcribed and nontranscribed genes, we defined genes with an average expression > 0 as transcribed and the remaining genes as nontranscribed in either control or CTE sample after QC and normalization described above. We counted mutation numbers in transcribed versus nontranscribed genes for each permutation and divided the observed counts over permuted counts to calculate an enrichment ratio for each transcriptional status in each clinical group. Then, a two-tailed Wilcoxon test was performed between transcribed and nontranscribed enrichment ratios in each group.

For chromatin accessibility enrichment analysis, we used the processed BED file of excitatory neurons (23) where the genome was first separated into nonoverlapping 1000-bp bins and snATAC-seq fragments were mapped to the bins. Similar to gene expression enrichment, the bins were ranked on the basis of their coverage and divided into equal-sized groups (10 groups for sSNVs, 5 groups for sIndels). Then we intersected both original and permuted mutation calls with genomic regions of each accessibility group. Enrichment ratios were calculated in the same way as above.

In addition to total somatic mutation enrichment, we performed signature-specific enrichment using the same NMF-based method for signature analysis above. For both original and permuted calls that overlapped with each expression or accessibility group, we fit the sSNV calls to signatures A and C and the sIndel calls to the double- and single-stranded signatures from META-CS. The observed/expected ratios were calculated as above.

Gene Ontology analysis

Gene Ontology enrichment analysis was performed on genes that overlapped with somatic mutations using GOseq (v.1.42.0) (58) after controlling for gene length bias. A null distribution was generated by

Wallenius approximation, and GO categories with less than 10 hits or more than 1000 genes were filtered out. P values of over- and underrepresented GO categories were adjusted for multiple testing using FDR. All GO categories with adjusted $P < 0.05$ in both CTE and neurotypical controls were reported in table S7.

REFERENCES AND NOTES

1. A. C. McKee *et al.*, The first NINDS/NIBIB consensus meeting to define neuropathological criteria for the diagnosis of chronic traumatic encephalopathy. *Acta Neuropathol.* **131**, 75–86 (2016). doi: 10.1007/s00401-015-1515-z; pmid: 26667418
2. A. C. McKee, T. D. Stein, P. T. Kiernan, V. E. Alvarez, The neuropathology of chronic traumatic encephalopathy. *Brain Pathol.* **25**, 350–364 (2015). doi: 10.1111/bpa.12248; pmid: 25904048
3. K. F. Bieniek *et al.*, The second NINDS/NIBIB consensus meeting to define neuropathological criteria for the diagnosis of chronic traumatic encephalopathy. *J. Neuropathol. Exp. Neurol.* **80**, 210–219 (2021). doi: 10.1093/jnen/nlab001; pmid: 33611507
4. R. C. Turner, B. P. Lucke-Wold, M. J. Robson, J. M. Lee, J. E. Bailes, Alzheimer's disease and chronic traumatic encephalopathy: Distinct but possibly overlapping disease entities. *Brain Inj.* **30**, 1279–1292 (2016). doi: 10.1080/02699052.2016.1193631; pmid: 27715315
5. R. A. Stern *et al.*, Clinical presentation of chronic traumatic encephalopathy. *Neurology* **81**, 1122–1129 (2013). doi: 10.1212/WNL.0b013e3182a55f7f; pmid: 23966253
6. A. C. McKee *et al.*, The spectrum of disease in chronic traumatic encephalopathy. *Brain* **136**, 43–64 (2013). doi: 10.1093/brain/aws307; pmid: 23208308
7. B. Falcon *et al.*, Novel tau filament fold in chronic traumatic encephalopathy encloses hydrophobic molecules. *Nature* **568**, 420–423 (2019). doi: 10.1038/s41586-019-1026-5; pmid: 30894745
8. Y. Shi *et al.*, Structure-based classification of tauopathies. *Nature* **598**, 359–363 (2021). doi: 10.1038/s41586-021-03911-7; pmid: 34588692
9. P. H. Montenigro, C. Bernick, R. C. Cantu, Clinical features of repetitive traumatic brain injury and chronic traumatic encephalopathy. *Brain Pathol.* **25**, 304–317 (2015). doi: 10.1111/bpa.12250; pmid: 25904046
10. B. Vogelstein *et al.*, Cancer genome landscapes. *Science* **339**, 1546–1558 (2013). doi: 10.1126/science.1235122; pmid: 23539594
11. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature* **458**, 719–724 (2009). doi: 10.1038/nature07943; pmid: 19360079
12. L. A. Garraway, E. S. Lander, Lessons from the cancer genome. *Cell* **153**, 17–37 (2013). doi: 10.1016/j.cell.2013.03.002; pmid: 23540688
13. I. Martincorena *et al.*, Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018). doi: 10.1126/science.aau3879; pmid: 30337457
14. I. Martincorena *et al.*, High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015). doi: 10.1126/science.aaa6806; pmid: 25999502
15. G. Genovese *et al.*, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014). doi: 10.1056/NEJMoa1409405; pmid: 25426838
16. F. Abascal *et al.*, Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021). doi: 10.1038/s41586-021-03477-4; pmid: 33911282
17. 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). doi: 10.1038/s41588-022-01180-2; pmid: 36163278
18. M. B. Miller *et al.*, Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714–722 (2022). doi: 10.1038/s41586-022-04640-1; pmid: 35444284
19. 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**, e2013106118 (2021). doi: 10.1073/pnas.2013106118; pmid: 33593904
20. M. H. Liu *et al.*, DNA mismatch and damage patterns revealed by single-molecule sequencing. *Nature* **630**, 752–761 (2024). doi: 10.1038/s41586-024-07532-8; pmid: 38867045
21. M. A. Lodato *et al.*, Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018). doi: 10.1126/science.aao4426; pmid: 29217584
22. V. Gonzalez-Pena *et al.*, Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024176118 (2021). doi: 10.1073/pnas.2024176118; pmid: 34099548
23. J. Ganz *et al.*, Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. *Cell* **187**, 1955–1970.e23 (2024). doi: 10.1016/j.cell.2024.02.025; pmid: 38503282
24. J. Mez *et al.*, Duration of American football play and chronic traumatic encephalopathy. *Ann. Neurol.* **87**, 116–131 (2020). doi: 10.1002/ana.25611; pmid: 31589352
25. M. L. Alosco *et al.*, Characterizing tau deposition in chronic traumatic encephalopathy (CTE): Utility of the McKee CTE staging scheme. *Acta Neuropathol.* **140**, 495–512 (2020). doi: 10.1007/s00401-020-02197-9; pmid: 32778942
26. A. C. McKee *et al.*, Neuropathologic and clinical findings in young contact sport athletes exposed to repetitive head impacts. *JAMA Neurol.* **80**, 1037–1050 (2023). doi: 10.1001/jamaneurol.2023.2907; pmid: 37639244
27. Z. Sondka *et al.*, COSMIC: A curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024). doi: 10.1093/nar/gkad986; pmid: 38183204
28. M. Jager *et al.*, Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res.* **29**, 1067–1077 (2019). doi: 10.1101/gr.246223.118; pmid: 31221724
29. A. R. Poetsch, The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput. Struct. Biotechnol. J.* **18**, 207–219 (2020). doi: 10.1016/j.csbj.2019.12.013; pmid: 31993111
30. F. Blokzijl, R. Janssen, R. van Boxtel, E. Cuppen, MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018). doi: 10.1186/s13073-018-0539-0; pmid: 29695279
31. I. T. Agaku *et al.*, Combustible and smokeless tobacco use among high school athletes—United States, 2001–2013. *MMWR Morb. Mortal. Wkly. Rep.* **64**, 935–939 (2015). doi: 10.15585/mmwr.mm6434a2; pmid: 26334565
32. F. Chagué *et al.*, Smokeless tobacco, sport and the heart. *Arch. Cardiovasc. Dis.* **108**, 75–83 (2015). doi: 10.1016/j.acvd.2014.10.003; pmid: 25497687
33. A. Serrano-Pozo, S. Das, B. T. Hyman, APOE and Alzheimer's disease: Advances in genetics, pathophysiology, and therapeutic approaches. *Lancet Neurol.* **20**, 68–80 (2021). doi: 10.1016/S1474-4422(20)30412-9; pmid: 33340485
34. K. Atherton *et al.*, Association of APOE genotypes and chronic traumatic encephalopathy. *JAMA Neurol.* **79**, 787–796 (2022). doi: 10.1001/jamaneurol.2022.1634; pmid: 35759276
35. M. A. M. Reijns *et al.*, Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature* **602**, 623–631 (2022). doi: 10.1038/s41586-022-04403-y; pmid: 35140396
36. Z. Zhou *et al.*, Recurrent patterns of widespread neuronal genomic damage shared by major neurodegenerative disorders. bioRxiv 2025.03.03.641186 [Preprint] (2025); https://doi.org/10.1101/2025.03.03.641186.
37. R. Madabhushi *et al.*, Activity-induced DNA breaks govern the expression of neuronal early-response genes. *Cell* **161**, 1592–1605 (2015). doi: 10.1016/j.cell.2015.05.032; pmid: 26052046
38. W. Wu *et al.*, Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440–444 (2021). doi: 10.1038/s41586-021-03468-5; pmid: 33767446
39. B. Jin *et al.*, Neurons accumulate disease-specific somatic genomic changes across tau pathologic states in Alzheimer's disease. bioRxiv 2025.05.26.656152 [Preprint] (2025); https://doi.org/10.1101/2025.05.26.656152.
40. A. C. McKee *et al.*, Chronic traumatic encephalopathy in athletes: Progressive tauopathy after repetitive head injury. *J. Neuropathol. Exp. Neurol.* **68**, 709–735 (2009). doi: 10.1097/NEN.0b013e3181a9d503; pmid: 19535999
41. J. D. Cherry, K. J. Babcock, L. E. Goldstein, Repetitive head trauma induces chronic traumatic encephalopathy by multiple mechanisms. *Semin. Neurol.* **40**, 430–438 (2020). doi: 10.1055/s-0040-1713620; pmid: 32674181
42. A. Fesharaki-Zadeh, Chronic traumatic encephalopathy: A brief overview. *Front. Neurol.* **10**, 713 (2019). doi: 10.3389/fneur.2019.00713; pmid: 31333567
43. G. D. Evrony *et al.*, Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012). doi: 10.1016/j.cell.2012.09.035; pmid: 23101622
44. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). doi: 10.1093/bioinformatics/btp324; pmid: 19451168
45. A. McKenna *et al.*, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010). doi: 10.1101/gr.107524.110; pmid: 20644199
46. X. Cai *et al.*, Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014). doi: 10.1016/j.celrep.2014.07.043; pmid: 25159146
47. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).
48. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi: 10.1093/bioinformatics/bty191; pmid: 29750242
49. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi: 10.1038/s41586-020-2308-7; pmid: 32461654
50. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015). doi: 10.18637/jss.v067.i01
51. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017). doi: 10.18637/jss.v082.i13
52. G. X. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017). doi: 10.1038/ncomms14049; pmid: 28091601
53. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021). doi: 10.1016/j.cell.2021.04.048; pmid: 34062119

54. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008). doi: 10.1088/1742-5468/2008/10/P10008

55. H. Mathys *et al.*, Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019). doi: 10.1038/s41586-019-1195-2; pmid: 31042697

56. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010). doi: 10.1093/nar/gkq603; pmid: 20601685

57. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016). doi: 10.1093/nar/gkv1189; pmid: 26553804

58. M. D. Young, M. J. Wakefield, G. K. Smyth, A. Oshlack, Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **11**, R14 (2010). doi: 10.1186/gb-2010-11-2-r14; pmid: 20132535

59. G. Dong, gldong/CTE_manuscript_code: v1.0, Zenodo (2025); https://doi.org/10.5281/zenodo.15300743.

ACKNOWLEDGMENTS

We thank R. Mathieu and T. Berisha from the Boston Children's Hospital Flow Cytometry Core, R. Krishnan from the Brigham and Women's Hospital Center for Neurologic Diseases Flow Cytometry Core Facility, and R. S. Hill and J. E. Neil for their administrative help. We thank the donors of postmortem tissues for their invaluable contributions to the advancement of science. **Funding:** National Institutes of Health (NIH) R56 AG079857 to A.Y.H., C.A.W., and E.A.L.; NIH R01 AG088082 to A.Y.H.; Alzheimer's Association Research Fellowship to A.Y.H.; NIH K08 AG065502 to M.B.M.; NIH DP2 AG086138 to M.B.M.; NIH R01 AG082346 to M.B.M.; Alzheimer's Disease Research program of the BrightFocus Foundation A20201292F to M.B.M.; Doris Duke Charitable Foundation Clinical Scientist Development Award 2021183 to M.B.M.; Suh Kyungbae Foundation to E.A.L.; NIH DP2 AG072437 to E.A.L.; NIH R01 NS032457-20S1 to C.A.W.; NIH R01 AG070921 and AG078929 to C.A.W. and E.A.L.; Templeton Foundation to C.A.W.; Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation to C.A.W. and E.A.L.; Howard Hughes Medical Institute to C.A.W.; National Institute of Neurological Disorders and Stroke U01NS086659, U01NS093334, U54NS115266, R01NS078337, R56NS078337, and K23NS102399 to A.C.M.; National Institute of Aging P30AG13846 to A.C.M.; Department of Veterans Affairs I01 CX001135 to A.C.M.; National Operating Committee on Standards for Athletic Equipment to A.C.M.; Nick and Lynn Buoniconti Foundation to A.C.M.; Concussion Legacy Foundation to A.C.M.; Andlinger Foundation to A.C.M.; WWE to A.C.M.; NFL to A.C.M. **Author contributions:** M.B.M., A.C.M., and C.A.W. conceived of the study, which was further developed by C.C.M. C.C.M. and M.B.M. performed tissue dissection, single-nuclei sorting, PTA genome amplification, and oversaw sequencing, with assistance from S.M.N., S.L.K., and G.A.M. K.S.-M.B., S.M.N., and S.P.W. performed META-CS genome amplification, with assistance from C.C.M. and M.B.M. and advice and interpretation from D.D.S. G.D. performed data analysis, with assistance from S.M. and J.K. G.D. developed and applied the software for mutation calling in META-CS data. A.C.M., M.U., J.D.C., and E.S. provided brain samples and clinicopathological data and assisted with data interpretation. G.D., C.C.M., S.M., and K.S.-M.B. prepared display items. G.D. and C.C.M. wrote the initial draft of the manuscript, with assistance from M.B.M., A.Y.H., S.M., and K.S.-M.B. M.B.M., A.Y.H., and E.A.L. oversaw experiments and analyses. C.A.W., E.A.L., M.B.M., and A.Y.H. jointly supervised the study and edited the manuscript. **Competing interests:** C.A.W. is a paid consultant (cash, no equity) to Third Rock Ventures and Flagship Pioneering and is on the clinical advisory board (cash and equity) of Maze Therapeutics. E.A.L. is on the scientific advisory board (cash, no equity) of Inocras. No research support was received from the four aforementioned companies. These companies did not fund and had no role in the conception or execution of this research project. G.D., C.C.M., A.Y.H., M.B.M., E.A.L., and C.A.W. have filed a preliminary patent application around findings reported in this paper. All other authors have no competing interests to declare. **Data and materials availability:** Sequencing data generated in this study have been deposited in the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) under accession number NG00183, with controlled use conditions set by human privacy regulations. Previously published PTA data for neurotypical controls can be downloaded from the database of Genotypes and Phenotypes (dbGaP) (accession number phs001485.v3.p1), and PTA data for AD can be downloaded from NIAGADS (accession number NG00121). All code used in this study is publicly available in Zenodo (*59*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adu1351

Figs. S1 to S17; Tables S1 to S7; MDAR Reproducibility Checklist

Submitted 24 October 2024; accepted 15 September 2025

10.1126/science.adu1351