Somatic Mosaicism in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Reveals Widespread Degeneration from Focal Mutations

Zinan Zhou[1,2,3,12], Junho Kim[1,2,3,4,12], August Yue Huang[1,2,3,12], Matthew Nolan[5], Junseok Park[1,2,3], Ryan Doan[1,3], Taehwan Shin[1,2,3], Michael B. Miller[1,6], Brian Chhouk[1,2,3], Katherine Morillo[1,2,3], Rebecca C. Yeh[1,2,3], Connor Kenny[1,2,3], Jennifer E. Neil[1,2,3,11], Chao-Zong Lee[5], Takuya Ohkubo[7,8], John Ravits[8], Olaf Ansorge[9], Lyle W. Ostrow[10], Clotilde Lagier-Tourenne[5,13], Eunjung Alice Lee[1,2,3,13] and Christopher A. Walsh[1,2,3,11,13]

Affiliations:
1. Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA.
2. Manton Center for Orphan Disease, Boston Children's Hospital, Boston, MA, USA.
3. Department of Pediatrics, Harvard Medical School, Boston, MA, USA .
4. Department of Biological Sciences, Sungkyunkwan University, Suwon, South Korea.
5. Department of Neurology, The Sean M. Healey and AMG Center for ALS at Mass General, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
6. Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
7. Department of Neurology, Yokohama City Minato Red Cross Hospital, Yokohama, Kanagawa, Japan.
8. Department of Neurosciences, School of Medicine, University of California at San Diego, La Jolla, CA, USA.
9. Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, Oxfordshire, UK.
10. Department of Neurology, Lewis Katz School of Medicine at Temple University, Philadelphia, USA.
11. Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA.
12. These authors contributed equally: Zinan Zhou, Junho Kim, August Yue Huang.
13. These authors jointly supervised this work: Clotilde Lagier-Tourenne, Eunjung Alice Lee, Christopher A. Walsh. Email: clagier-tourenne@mgh.harvard.edu; ealee@childrens.harvard.edu; christopher.walsh@childrens.harvard.edu.

Abstract

Although mutations in dozens of genes have been implicated in familial forms of amyotrophic lateral sclerosis (fALS) and frontotemporal degeneration (fFTD), most cases of these conditions are sporadic (sALS and sFTD), with no family history, and their etiology remains obscure. We tested the hypothesis that somatic mosaic mutations, present in some but not all cells, might contribute in these cases, by performing ultra-deep, targeted sequencing of 88 genes associated with neurodegenerative diseases in postmortem brain and spinal cord samples from 404 individuals with sALS or sFTD and 144 controls. Known pathogenic germline mutations were found in 20.6% of ALS, and 26.5% of FTD cases. Predicted pathogenic somatic mutations in ALS/FTD genes were observed in 2.7% of sALS and sFTD cases that did not carry known pathogenic or novel germline mutations. Somatic mutations showed low variant allele fraction (typically <2%) and were often restricted to the region of initial discovery, preventing detection through genetic screening in peripheral tissues. Damaging somatic mutations were preferentially enriched in primary motor cortex of sALS and prefrontal cortex of sFTD, mirroring regions most severely affected in each disease. Somatic mutation analysis of bulk RNA-seq data from brain and spinal cord from an additional 143 sALS cases and 23 controls confirmed an overall enrichment of somatic mutations in sALS. Two adult sALS cases were identified bearing pathogenic somatic mutations in *DYNC1H1* and *LMNA,* two genes associated with pediatric motor neuron degeneration. Our study suggests that somatic mutations in fALS/fFTD genes, and in genes associated with more severe diseases in the germline state, contribute to sALS and sFTD, and that mosaic mutations in a small fraction of cells in focal regions of the nervous system can ultimately result in widespread degeneration.

Introduction

Amyotrophic lateral sclerosis (ALS), a disease in which premature loss of upper and lower motor neurons (UMNs and LMNs) leads to fatal paralysis, shows clinical, genetic, and pathological overlap with frontotemporal dementia (FTD), a neurodegenerative disorder characterized by behavioral, language, and memory dysfunction[1]. 5-22% of individuals with ALS develop FTD, and ≈ 15% of those with FTD eventually develop ALS[2]. ALS and FTD also share common pathology, with cytoplasmic inclusions of TAR DNA binding protein (TDP-43) found in almost all ALS brains and in half of FTD brains[3,4]. FTD brains lacking TDP-43 inclusions mainly show tau pathology. ALS typically begins focally and spreads regionally as the disease progresses[5,6], although whether degeneration begins in UMNs, LMNs, or both simultaneously, has remained controversial[7,8], with some studies suggesting that focality can manifest independently in UMNs and LMNs[5,9]. TDP-43 pathology also follows stereotypical patterns in ALS and FTD brains[9-11], thought to reflect focal onset and intercellular transmission of TDP-43 inclusions in a prion-like manner, as shown in cell and animal models[12-18].

Whereas over 30 genes are implicated in ALS and FTD[19], most causative genes are linked to familial ALS (fALS) and FTD (fFTD), while 90-95% of cases are sporadic ALS (sALS) and FTD (sFTD) without a family history[20]. Prospective studies of ALS revealed a higher number of cases stemming from a genetic basis, regardless of whether a family history is documented[21], with the underestimation of genetic cases probably reflecting multiple factors, including incomplete ascertainment, death from other causes before diagnosis, and incomplete disease penetrance. Therefore, genetic screening of ALS/FTD genes is needed to fully examine the contribution of germline mutations in sporadic cases.

The focal onset of ALS and FTD, their stereotypical spread, and the increased risk in smokers[22], have raised interest in potential roles of somatic mosaic mutations in the pathogenesis of ALS and FTD[23]. Somatic mutations are increasingly recognized as prevalent in normal-appearing tissues, but somatic mutations responsible for neurological conditions are often limited to the central nervous system (CNS)[24], and hence undetectable through DNA sequencing of non-CNS tissues. Recent studies have evaluated the contributions of somatic mutation to Alzheimer's and Parkinson's disease directly using postmortem brain tissues[25].

In this study, we assessed potential contributions of germline and somatic mutations — distinguished by their variant allele frequencies (VAFs) — to sALS and sFTD using ultra-deep sequencing of a panel of neurodegeneration-associated genes on postmortem tissues of various brain regions and spinal cords from >400 unique sALS and sFTD cases. Our study revealed that pathogenic germline mutations are more common than previously appreciated in sALS and sFTD cases, supporting the underestimation of ALS and FTD cases with underlying genetic causes. In addition, we identified novel predicted pathogenic somatic mutations in 2.7% of the sALS and sFTD cases without known or novel pathogenic germline mutations. Protein-altering (missense/nonsense/frameshift) somatic mutations showed significant enrichment in sALS and sFTD cases and in disease-affected brain regions, supporting roles in disease pathogenesis. Regional analysis revealed focality of predicted pathogenic somatic mutations in primary motor cortex and spinal cord, supporting independent disease initiation in UMNs and LMNs, but also strongly supporting models of ALS and FTD in which the disease spreads beyond a relatively confined region containing a somatic mutation.

Results

Ultra-deep targeted sequencing of neurodegenerative genes in sALS and sFTD brains

To directly detect somatic mutations in sALS and sFTD brains, we obtained post-mortem frozen tissues of several brain regions and spinal cords from individuals diagnosed with sALS or sFTD, as well as from age-matched controls through the Massachusetts Alzheimer's Disease Research Center, Oxford Brain Bank, and Target ALS Foundation (Fig. 1a and Supplementary Table 1). Additional brain tissues from ALS, FTD and control cases, without a record of family history but with an age of death above 45 years old, were also obtained from the NIH NeuroBioBank. We designed a molecular inversion probe (MIP) panel targeting the exons and

106 exon-intron junctions of 88 neurodegeneration-related genes[26], which included 34 ALS/FTD
107 genes, 10 Alzheimer's disease genes, 28 Parkinson's disease genes, and 16 genes associated with
108 other rare neurodegenerative disorders (Supplementary Table 2). We performed MIP panel
109 sequencing at ~1,800X average sequencing depth (Fig. 1b and Extended Data Fig. 1), with a
110 similar distribution of sequencing depth across batches, disease conditions, and tissue regions
111 (Extended Data Fig. 1). The variance of depth, along with the batch and sample information,
112 were considered as factors in the mutation burden test . A total of 937, 364, and 516 samples
113 from 291 ALS, 117 FTD, and 144 neurotypical control individuals respectively were sequenced
114 (Fig. 1a, 1c and Supplementary Table 1). Of the ALS and FTD cases, 8 were diagnosed with
115 both ALS and FTD, and were therefore counted for each condition.
116
117 **Pathogenic germline mutations in sALS and sFTD cases**
118 We first identified pathogenic germline single-nucleotide variants (SNVs) and short
119 insertions and deletions (indels) using GATK followed by multiple variant filters (Fig. 1d). The
120 functional impact and predicted pathogenicity of identified germline mutations were annotated
121 using ANNOVAR[27] and multiple clinical databases. In addition, the most common inherited
122 cause of ALS and FTD, a hexanucleotide repeat expansion in the *C9ORF72* gene[28,29], was
123 genotyped by a repeat-primed PCR assay (Extended Data Fig. 2). Overall, 20.6% (60/291) of
124 ALS, 26.5% (31/117) of FTD and 0.7% (1/144) of control cases showed *C9ORF72* repeat
125 expansions or pathogenic germline mutations in ALS and FTD genes that have been previously
126 reported (Fig. 2a, Supplementary Table 3, 4). Known and novel missense mutations in ALS/FTD
127 genes represented the most prevalent mutation type (Fig. 2b). *C9ORF72* repeat expansion was
128 the most frequently mutated gene followed by known and novel pathogenic germline mutations
129 in *SOD1* for ALS, and *GRN* and *MAPT* mutations for FTD cases (Fig. 2c and 2d). The overall
130 fractions of *C9ORF72* repeat expansion carriers — 10.6% for ALS-only cases and 12.0% for
131 FTD-only cases — slightly exceeded those reported in previous studies, yet they remained
132 notably lower than the rates observed in fALS and fFTD cases[30-32]. Three carriers of the
133 *C9ORF72* repeat expansion also had known pathogenic mutations in other genes associated with
134 ALS/FTD (Fig. 2d and Supplementary Table 3), aligning with previous studies that have
135 demonstrated instances of oligogenic inheritance involving *C9ORF72* repeat expansions and
136 other pathogenic mutations in certain sALS and sFTD cases[33,34].
137 Our pathogenicity prediction found pathogenic germline mutations in dominant
138 ALS/FTD genes besides *C9ORF72* repeat expansions in 14.1% of ALS, 19.7% of FTD, and
139 5.6% of control cases (Fig. 2a, Supplementary Table 3, 4). The odds ratios for the presence of
140 pathogenic mutations in ALS and FTD cases, compared to control cases, were 2.78 (95% CI:
141 1.24-7.07, $p=9.3\times10^{-3}$) and 4.14 (95% CI: 1.70-11.17, $p=8.2\times10^{-4}$) respectively, suggesting
142 pathogenic mutations are enriched in both ALS and FTD cases. Not surprisingly, all previously
143 reported pathogenic mutations were predicted to be pathogenic. Most novel pathogenic
144 mutations were nonsynonymous SNVs that would require experimental validation to confirm
145 their functional impact. However, two novel *GRN* frameshift mutations (p.L46Rfs*18 and
146 p.D250Tfs*6) identified in FTD cases are probably disease-causing (Supplementary Table 3),
147 since loss-of-function *GRN* mutations are known to cause FTD in a dominant manner[35,36].
148 When we considered previously unreported but likely pathogenic germline mutations,
149 another 12 disease cases exhibited potential instances of oligogenic inheritance (Fig. 2d). Of
150 these, five individuals carried *C9ORF72* repeat expansions alongside novel pathogenic germline
151 mutations in other ALS/FTD genes, while another five cases had known pathogenic germline

152 mutations in *GRN*, *SOD1*, and *MAPT* genes, in combination with novel predicted pathogenic
153 germline mutations in other ALS/FTD genes. Two patients carried multiple novel pathogenic
154 germline mutations. These findings provide additional evidence for oligogenic inheritance of
155 ALS and FTD[33,34,37,38] (Fig. 2d). We also found 13 FTD cases to have germline mutations in
156 genes previously linked to ALS only (*NEK1*, *SETX*, *ATP13A2*, *ALS2*, *ANXA11*, *DCTN1, FIG4*
157 *and VAPB*) and one ALS case to have a predicted pathogenic missense mutation in the FTD-
158 associated *MAPT* gene (Fig. 2d). These crossover mutations between ALS and FTD reinforce the
159 overlap between both diseases from shared underlying mechanisms.
160
161 **Identification of somatic SNVs and indels from MIP sequencing data**
162 We developed a custom pipeline integrating RePlow[39], Mutect2[40], and Pisces[41] for calling
163 somatic SNVs and indels in our MIP sequencing data (Fig. 1d). We selected somatic mutations
164 identified by at least two of the three callers (double-called mutations) followed by multi-step
165 variant filters to remove false positive candidates. Unlike heterozygous germline mutations with
166 variant allele fractions (VAFs) around 50%, heterozygous somatic mutations have VAFs less
167 than 50%, and we only called somatic mutations with VAFs below 40%. To benchmark our
168 pipeline, we performed a spike-in experiment by mixing two human samples from the Genome
169 in a Bottle Consortium (GIAB) at variant allele fractions (VAFs) of 10%, 5%, 2.5%, 1%, and
170 0.5% (Extended Data Fig. 3a). Double-called mutations identified by Mutect2 and Pisces were
171 excluded from the final call set due to high false positive and low validation rates (Extended
172 Data Fig. 3b, c). High sensitivity and precision were achieved for the remaining Replow-based
173 double-called mutations (Replow-Mutect2 and Replow-Pisces) while maintaining a low false
174 positive rate across the low VAFs compared to the somatic mutations called by each caller. The
175 MIP sequencing and our custom pipeline together allowed us to confidently identify somatic
176 mutations with a low false positive rate at VAF as low as 0.5%. The observed VAFs of somatic
177 mutations were well in line with the target VAFs at all five VAF levels (Extended Data Fig. 3).
178 The custom pipeline identified 167 somatic SNVs and indels from our MIP sequencing
179 data (Supplementary Table 5). The VAF distribution of identified somatic mutations was similar
180 between disease and control cases at high VAF levels (>5%), but low-VAF mutations were more
181 common in disease cases (Extended Data Fig. 4). Forty-one somatic candidates were selected for
182 validation and 87.8% of them were confirmed by deep amplicon sequencing (Supplementary
183 Table 6). The VAFs of validated candidates in amplicon sequencing showed a strong correlation
184 with their original VAFs in the MIP sequencing data (Fig. 3a).
185
186 **Somatic mutations in disease-relevant genes are enriched in ALS and FTD cases lacking**
187 **pathogenic germline mutations**
188 To examine the burden and potential roles of somatic mutations in ALS and FTD, we
189 focused on cases that lacked known or novel pathogenic germline mutations (referred to as
190 germline-free cases). Ninety-five unique somatic mutations in neurodegeneration-related genes
191 were identified in 696, 243, and 516 samples from 216 ALS germline-free cases, 76 FTD
192 germline-free cases, and 144 neurotypical controls, respectively. Most somatic mutations (80%,
193 76 out of 95 unique mutations) were focal, identified only in one tissue region of an individual
194 (Fig. 3b), and at very low VAFs (Extended Data Fig. 4), suggesting that they likely arose after
195 gastrulation[42], and are likely to have been confined to nervous tissue. Mutational signature
196 analysis using Mutalisk[43] demonstrated that clock-like signatures (SBS5 and SBS1) were the
197 predominant mutational signatures (Extended Data Fig. 5). Recent work has identified their

198 presence in brain development[44,45], and SBS1 reflects deamination of methylated cytosine during
199 cell division and mitosis.
200 Our MIP panel contained not only ALS/FTD genes but also genes involved in other
201 dementia. We first focused on somatic mutations in all the neurodegenerative genes. For the
202 somatically mutated genes, there was a clear separation between the disease and control groups
203 (Fig. 3c). Indeed, just one protein-altering somatic mutation was observed among all controls,
204 while 15 and 7 were observed in ALS and FTD cases, respectively. These protein-altering
205 somatic mutations were significantly enriched in ALS and FTD cases (Fig. 4a; p=0.013 and
206 p=0.011) when tested using a linear mixed-effect regression model, which considers multiple
207 potential confounding factors, suggesting that some or all of them were potentially disease-
208 causing.
209 The enrichments of somatic mutations in neurodegenerative genes showed striking
210 topographic patterns, with exonic somatic mutations showing enrichment exclusively in disease-
211 affected tissue regions for both FTD and ALS germline-free cases. The prefrontal cortex showed
212 enrichment for somatic FTD mutations, and the primary motor cortex for ALS (Fig. 4b), while
213 the premotor cortex—located immediately in between these two regions—showed no enrichment
214 for either condition, as was the case for other tested cerebral cortical regions as well. The spinal
215 cord in ALS had only a modest increase in protein-altering somatic mutations, although this
216 analysis is limited by a small number of control spinal cord samples and resultant wide
217 confidence intervals (Fig. 4b). For the prefrontal cortex of FTD and the primary motor cortex of
218 ALS, enrichments of protein-altering somatic mutations in germline-free cases were even more
219 significant than the overall enrichments of exonic somatic mutations (Fig. 4b; p=0.043 and
220 $p=9.1\times10^{-3}$, $p=6.8\times10^{-3}$ and $p=2.4\times10^{-3}$ for exonic and protein-altering mutations in ALS and
221 FTD germline-free cases, respectively; linear mixed model), further supporting the pathogenic
222 roles of the identified somatic mutations.
223 We further assessed somatic mutations in genes specifically related to ALS and FTD and
224 found that somatic mutations in each were enriched in genes relevant to that corresponding
225 condition. Exonic and protein-altering mutations were specifically enriched in ALS-related genes
226 in germline-free ALS samples (Fig. 4c; p=0.029 and p=0.017 for exonic and protein-altering
227 mutations, linear mixed model). Moderate enrichments were observed for exonic and protein-
228 altering mutations in FTD-related genes in germline-free FTD samples. In fact, less than half of
229 FTD cases have pathological TDP-43 protein aggregates, while the other half have Tau protein
230 aggregates[4]. We thus checked the contribution of Tau proteinopathy-related genes, including
231 genes associated with Alzheimer's disease (AD), together with FTD-related genes and found
232 nominally significant enrichment of exonic and protein-altering somatic mutations only in
233 germline-free FTD cases (Fig. 4c; p=0.046 for both exonic and protein-altering mutations, linear
234 mixed model). Our FTD cases could not be categorized into those related to TDP-43 or Tau
235 proteinopathies due to the lack of relevant pathological information, hindering our ability to
236 examine the potential enrichment of somatic mutations within these distinct categories. On the
237 other hand, no protein-altering mutation was observed in any of the ALS/FTD genes in control
238 samples (Fig. 3c). The exclusive and diagnosis-specific enrichments of functional somatic
239 mutations suggest that most or all somatic mutations contribute to the pathogenesis of sALS and
240 sFTD.
241
242

243 **Pathogenic somatic mutations have restricted regional distributions and are enriched in**
244 **hypodiploid cells**
245 Pathogenicity prediction of somatic mutations resulted in 8 predicted pathogenic somatic
246 SNVs in previously known ALS and FTD/Tau-proteinopathy genes (Supplementary Table 7),
247 which account for 3.2% and 2.6% of germline-free ALS and FTD cases, respectively (2.7% for
248 all the germline-free sALS and sFTD cases). All mutations in ALS cases were observed in
249 primary motor cortex or spinal cord, the most severely affected regions in ALS, emphasizing the
250 remarkable topographic specificity of the mutations. In addition, a predicted pathogenic somatic
251 SNV in *APP* (p.R328Q) was identified in primary motor cortex of a sporadic case that showed
252 both ALS and FTD. All somatic mutations occurred in disease genes with dominant inheritance
253 when found in the germline setting, except for one sALS case with a somatic *ALS2* (p.T787R)
254 mutation identified in spinal cord. *ALS2* is an autosomal recessive disease gene[46,47], and the same
255 individual carried an *ALS2* (p.Q24R) germline mutation in addition to the identified somatic
256 mutation. Both *ALS2* mutations were predicted to be pathogenic, suggesting that they initiate
257 disease in a "second hit" autosomal recessive manner at the cellular level in a small proportion of
258 cells in the spinal cord and again further supporting the likely contribution of somatic variants to
259 pathogenesis.
260 We selected four predicted pathogenic somatic SNVs in ALS/FTD genes-- *TIA1*
261 (p.H54R), *MATR3* (p.K594I), *ALS2* (p.T787R), and *TARDBP* (p.L248F), and the predicted
262 pathogenic *APP* somatic SNV (p.R328Q)--to study in greater detail in terms of regional and cell-
263 type distributions. Amplicon sequencing across multiple brain and spinal cord regions showed
264 that three of the five somatic SNVs [*MATR3* (p.K594I), *APP* (p.R328Q), *TARDBP* (p.L248F)]
265 were restricted to the primary motor cortex (Fig. 5a and Supplementary Table 8). The other two
266 somatic SNVs [*TIA1* (p.H54R) and *ALS2* (p.T787R)] had the highest VAFs in the spinal cord
267 [2.16% for *TIA1* (p. H54R) and 0.97% for *ALS2* (p.T787R)], where they were originally
268 identified, and were also present in other brain regions at very low VAFs [0.15-1.05% for *TIA1*
269 (p.H54R), 0.16% - 0.59% for *ALS2* (p.T787R)] (Fig. 5a and Supplementary Table 8). All five
270 somatic SNVs were absent in cerebellum. The ultra-low levels and limited distribution of these
271 somatic SNVs suggest that they probably arose late in development and were thus likely
272 excluded from non-CNS tissues. Together with the enrichment of exonic and protein-altering
273 somatic mutations in disease-affected tissue regions, these findings also support the focal onset
274 of ALS at the genetic level in these somatic cases. Cells carrying damaging somatic mutations
275 could form initial lesions, likely TDP-43 inclusions, in UMNs and LMNs, but these must have
276 ultimately spread to other regions of the motor system that lacked or carried exceedingly low
277 levels of the mutation, but which nonetheless showed robust pathology post mortem otherwise
278 indistinguishable from germline cases.
279 We then determined the presence of these five somatic SNVs in different cell types by
280 performing amplicon sequencing of DNA from neuronal (NeuN+), glial (NeuN-), diploid,
281 polyploid, and hypodiploid nuclei isolated by fluorescence-activated nuclei sorting (FANS) from
282 the tissue regions in which they were originally identified (Extended Data Fig. 6). Interestingly,
283 *TIA1* (p.H54R), *MATR3* (p.K594I), and *ALS2* (p.T787R) mutations were enriched in hypodiploid
284 nuclei (Fig. 5b), which likely represent apoptotic cells with DNA fragmentation and cell
285 death[48,49]. Enrichment of these three mutations in hypodiploid cells indicates a possible role in
286 the pathogenic process, suggesting that they might be involved in inducing cell death.
287 Surprisingly, these three mutations were identified in all cell fractions, but were more enriched in
288 non-neuronal cells compared to neurons (Fig. 5b). This finding also implies that neurons may

289 exhibit a cell-type specific vulnerability to damaging somatic mutations in ALS/FTD genes. In
290 contrast, the depletion of the *APP* mutation from hypodiploid cells, and its relative enrichment in
291 non-neuronal cells compared to neurons (Fig. 5b), align with models proposing important actions
292 of AD risk genes in non-neuronal cells including microglia and astrocytes, potentially leading to
293 secondary neuronal loss[50]. However, further research is needed to confirm and better understand
294 these potential associations and mechanisms. The *TARDBP* (p.L248F) mutation was found in a
295 primary motor cortex sample with a very low VAF (≈ 0.5% upon validation). However, when
296 isolated cell fractions were tested, the mutation was not detected in any of them. This suggests
297 that the mutation was only present in the specific area where it was initially discovered and did
298 not extend to nearby regions. This conclusion was confirmed by amplicon sequencing of a
299 second tissue sample from the primary motor cortex, where it was also absent.
300
301 **RNA-MosaicHunter identifies additional pathogenic somatic mutations in bulk RNA-seq**
302 **data of sALS cases**
303 To complement our targeted sequencing of neurodegenerative genes, which identified
304 pathogenic somatic mutations in a small proportion of sALS and sFTD cases in known genes, we
305 performed a transcriptome-wide screen for somatic mutations using RNA-seq data to explore
306 whether genes not normally associated with these conditions might cause them in the mosaic
307 state. We profiled pathogenic somatic mutations in all expressed genes in bulk RNA-seq data
308 generated from 789 postmortem brain and spinal cord tissue samples of 143 sALS cases and 23
309 age-matched controls by the New York Genome Center ALS Consortium (Supplementary Table
310 9; 81 and 11 of the sALS and control cases respectively were included in our MIP sequencing)
311 using RNA-MosaicHunter, a tool capable of calling clonal somatic mutations from bulk RNA-
312 seq data with a Bayesian probabilistic model. Because of the limited coverage of bulk RNA-seq
313 data, RNA-MosaicHunter only has sensitivity to detect somatic mutations VAFs >≈5%, and
314 discards somatic mutations at ultra-low levels. We found significant increases in total somatic
315 mutations in sALS cases not carrying pathogenic germline mutations (Extended data Fig. 7;
316 p=0.007). Additionally, there was a higher burden of somatic mutations predicted to be
317 damaging in germline-free sALS cases; although this trend did not reach statistical significance
318 (Extended data Fig. 7; p=0.058). Overall, these findings further confirmed that somatic mutations
319 may contribute to the development of sALS.
320 Interestingly, somatic SNVs in *DYNC1H1* and *LMNA* were identified in multiple CNS
321 regions of two sALS cases that did not harbor other pathogenic germline or somatic mutations
322 (Fig. 6a and Supplementary Table 10, both cases were included in the MIP sequencing).
323 Heterozygously acting, generally de novo, mutations in *DYNC1H1* and *LMNA* have been found
324 in patients with phenotypes resembling spinal muscular atrophy (SMA)[51-54], a motor neuron
325 disease genetically distinct but sharing some pathological overlap with ALS, including loss of
326 lower motor neurons, denervation of neuromuscular junction, and muscle atrophy[55]. Analysis of
327 whole-genome sequence data of the two cases for germline mutations in *SMN1*, the most
328 commonly mutated genes in SMA, did not identify pathogenic germline mutations. Both
329 individuals carrying these somatic mutations had leg-onset ALS with TDP43 pathology
330 predominantly in spinal cord and to a lesser extent in motor cortex (Fig. 6a-c). We further
331 investigated their regional mutation distribution using amplicon sequencing. The *LMNA*
332 (p.H566Y) somatic mutation was detected in all the tested brain and spinal cord regions with
333 VAFs ranging from 5.3 to 12.3% (Fig. 6d and Supplementary Table 8). The *DYNC1H1*
334 (p.R1962C) somatic mutation was also detected in all the tested CNS regions with VAFs ranging

335 from 0.1% to 5.2%, but the VAFs of the mutation were extremely low in the cerebellum (0.1%),
336 thoracic spinal cord (0.8%) and lumbar spinal cord (0.8%) (Fig. 6d and Supplementary Table 8).
337 Notably, the *DYNC1H1* (p.R1962C) mutation was undetectable in cultured fibroblasts from the
338 patient (Supplementary Table 8), indicating that the mutation arose late in development and was
339 likely limited to the CNS. The broad distribution of these two somatic mutations aligns with our
340 previous finding that somatic mutations with more than 5% VAFs are typically detected
341 throughout the CNS[56], with the low levels of the mutation in lumbar spine potentially reflecting
342 death of the motor neurons carrying this mutation. The *DYNC1H1* p.R1962C mutation is known
343 to be highly pathogenic, as it completely abolishes the motor function of the dynein complex *in*
344 *vitro*[57], and germline *DYNC1H1* p.R1962C mutations have been found in patients with
345 malformations of cortical development and delayed psychomotor development[58,59]. Although the
346 *LMNA* (p.H566Y) mutation was not previously reported, *LMNA* mutations cause autosomal
347 dominant laminopathies including Hutchinson-Gilford progeria and congenital muscular
348 dystrophy, which are characterized by congenital defects and increased early lethality[60,61]. Thus,
349 both genes can cause lethal diseases with pediatric age of onset, which may ordinarily preclude
350 the appearance of ALS, but the mosaic state could allow for a normal early life and the onset of a
351 degenerative disorder later in life. These data suggest that further genome-wide exploration of
352 brain tissue for somatic mutations could reveal additional ALS genes that cause early lethality in
353 the germline state.
354
355 **Discussion**
356 Our data provide several important insights into sALS and sFTD. First, we found that
357 about 30% of both conditions show known or novel, likely pathogenic germline mutations in
358 ALS or FTD genes, which advocates for a shift from family history-based to genetic testing-
359 based classification of ALS and FTD cases. Second, we find that a small but important fraction
360 (~2.7%) of germline-free sporadic cases harbor predicted pathogenic somatic variants in known
361 ALS or FTD genes, with the distribution of these mutations being disease and brain region-
362 specific, providing proof of concept of a potentially important contribution of somatic mutations
363 to pathogenesis. Finally, we find examples of genes associated with severe pediatric degenerative
364 diseases that can be present in ALS in the somatic state, potentially broadening the spectrum of
365 causative genes for these conditions.
366 While the case-control enrichment of somatic variants suggests a role in pathogenesis,
367 these somatic variants are present at surprisingly low VAFs and with patterns of topographic
368 restriction that match disease onset. It is very likely that these pathogenic somatic mutations
369 arose at a late stage of development and were not shared by other tissue regions. In the most
370 extreme case, the *TARDBP* (p. L248F) somatic SNV was even undetectable in tissue adjacent to
371 the original sampling site. The nature of these focal somatic events would prevent them from
372 being identified through routine genetic testing with blood or other peripheral samples. The
373 focality of these mutations in the nervous system also suggests a mechanism by which
374 degeneration may spread from a site containing mutant cells to eventually cause loss of neurons
375 in regions that do not carry the mutation. This process is thought to involve the TDP-43
376 proteinopathy as supported by recent studies in cell and animal models[12-18]. Identification of
377 predicted pathogenic somatic mutations in the primary motor cortex and in spinal cord from
378 individuals with ALS suggests potential onset of disease in either UMNs or LMNs but eventual
379 involvement of both. Our cell-type analysis revealed that several predicted pathogenic somatic
380 mutations were more enriched in glia than neurons. However, the reduced abundance in neurons

381 might also reflect the loss of neurons carrying these somatic mutations. This was reinforced by
382 our observations that three out of the four tested somatic mutations were more prevalent in
383 hypodiploid cells, which likely represent apoptotic cells. The potential harm inflicted on neurons
384 by these mutations once again bolsters the concept of a focal onset of ALS. Neurons carrying
385 these mutations constitute the initial lesion and subsequently undergo cell death. The demise of
386 these neurons could further reduce their presence, leading to a reduction in the VAFs of the
387 mutations compared to their levels at the time of initial emergence.
388 Although only about 2.7% of germline-free ALS and FTD cases had predicted
389 pathogenic somatic mutations in our MIP sequencing data, this is likely greatly underestimated
390 because of the limited sensitivity of even our deep panel sequencing approach to detect somatic
391 mutations at ultra-low levels (Extended data Fig. 4). The detection of somatic mutations with low
392 VAFs remains a technical challenge[45], but improved duplex sequencing approaches promise the
393 ability in the future to systematically sample somatic mutations at virtually all allele frequency
394 levels. Given that somatic mutations at very low levels and in focal regions appear capable of
395 creating a spreading disease, it will require very deep analysis to determine the lower allele
396 frequency range of variants that is capable of initiating this process. Variant detection is also
397 limited by availability of samples from regions across the CNS.
398 Our identification of candidate somatic SNVs in *DYNC1H1* (p. R1962C) and *LMNA* (p.
399 H566Y) using RNA-seq analysis of sALS cases suggests that genes that predispose their carriers
400 to ALS and FTD by somatic mutations may include genes distinct from those discovered in
401 germline cases. Certain alleles in both *DYNC1H1* and *LMNA* are associated with motor neuron
402 degeneration in the form of SMA, so they are capable of predisposing to neuronal degeneration,
403 but also in both cases, other alleles (including the *DYNC1H1* p. R1962C allele[58,59]) cause severe
404 pediatric disease that would normally mask the possibility of late-life ALS. This result suggests
405 that a wider range of ALS genes and alleles could exist in the somatic state that cannot be
406 observable in the germline state due to their association with early-onset severe disease. This
407 raises an exciting prospect that future genome-wide approaches, such as deep whole-genome or
408 exome sequencing of a cohort of ALS cases, could shed light not only on additional somatic
409 genetic mechanisms and their contributions to ALS, but also on the topographic patterns of
410 spread of pathology from focal sites.
411

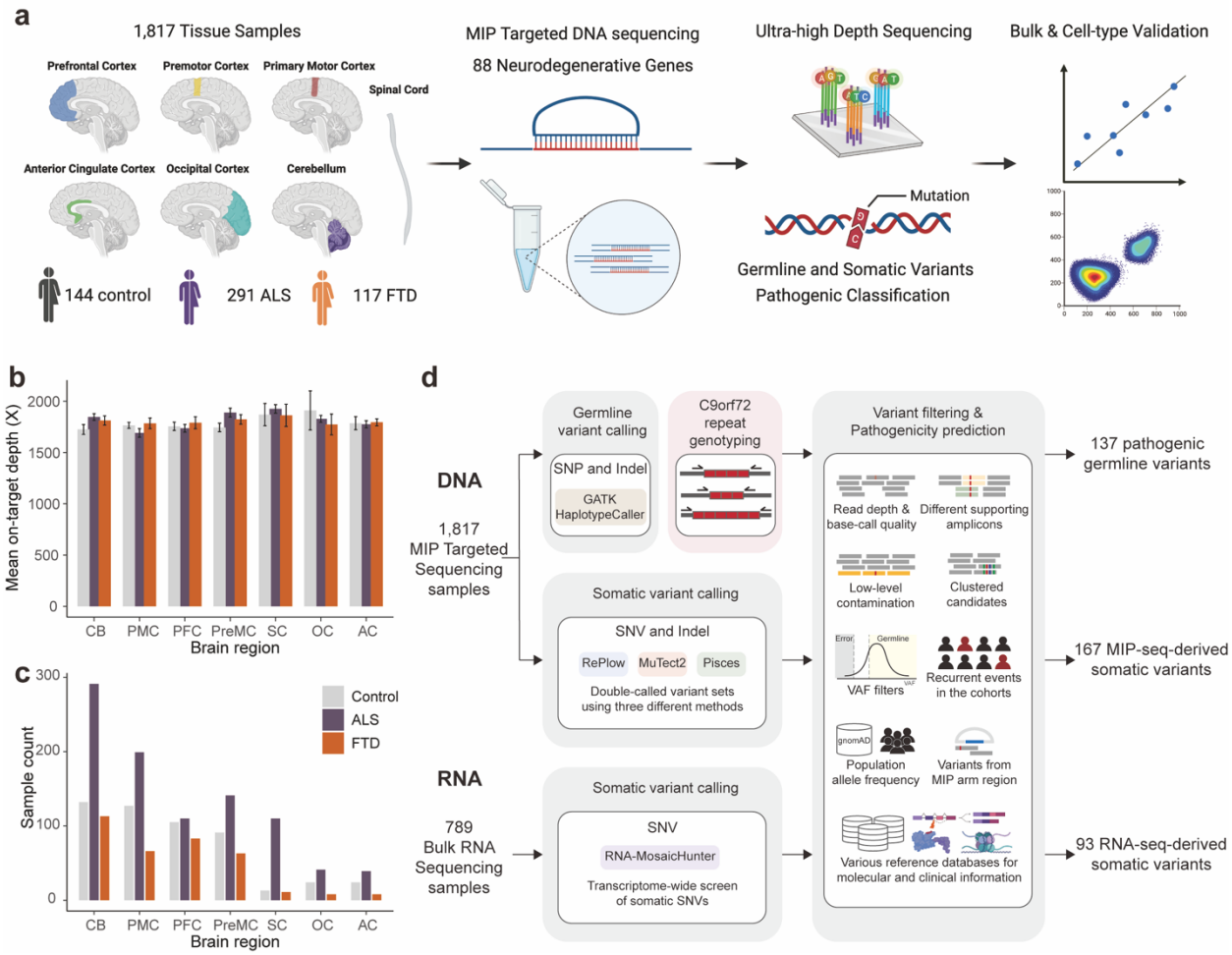


Fig: 1. Experimental and analysis strategies. (a) Overall scheme of the experiments. Genomic DNA isolated from 1,817 postmortem tissue samples of multiple brain regions and spinal cords of 144 control, 291 ALS, and 117 FTD cases were used for molecular inversion probe (MIP) capture sequencing with ultra-high depth. (b, c) Mean sequencing depth and number of tissue samples in different brain regions and spinal cords of control, ALS, and FTD cases. Control, n=516; ALS, n=937; FTD, n=364. CB: cerebellum; PMC: primary motor cortex; PFC: prefrontal cortex; PreMC: premotor cortex; SC: spinal cord; OC: occipital cortex; AC: anterior cingulate cortex. Error bars, 95% CI (d) Methodological pipelines to identify germline and somatic variants. Germline variants were called by GATK HaplotypeCaller. *C9ORF72* genotype of ALS and FTD cases were determined by repeat-primed PCR. Somatic variants were called by RePlow, MuTect2, and Pisces. Additional somatic variants were called from 789 bulk RNA-seq profiles of multiple brain regions and spinal cords of ALS cases generated by the New York Genome Center ALS Consortium using RNA-MosaicHunter.

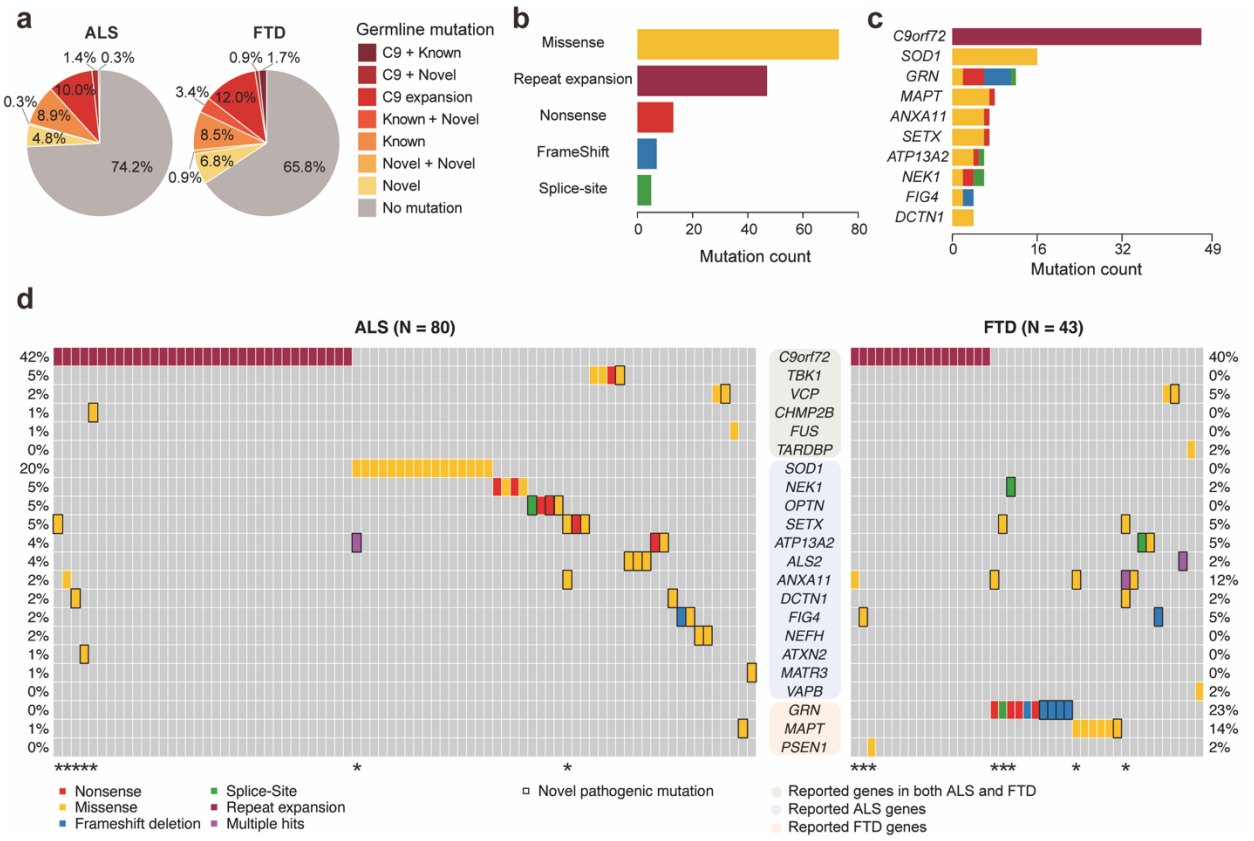


Fig. 2: *C9ORF72* repeat expansion and pathogenic germline variants in ALS/FTD genes are prevalent in ALS and FTD. (a) Proportions of ALS and FTD cases with *C9ORF72* repeat expansion, known, and novel pathogenic germline variants of ALS/FTD genes. Cases with multiple pathogenic mutations are indicated with '+' sign. (b) Distribution of *C9ORF72* repeat expansion and known and novel pathogenic germline variants in ALS/FTD genes classified by mutation types. (c) Ranking of the top 10 mutated ALS/FTD genes.(d) Visualization of ALS and FTD cases (vertical columns) with known and novel pathogenic germline variants (horizontal rows) in ALS/FTD genes. Color codes indicate the types of mutations. Rectangular outline represents novel variants. Genes are grouped by their known involvement in the diseases. * indicates cases with multiple pathogenic mutations.

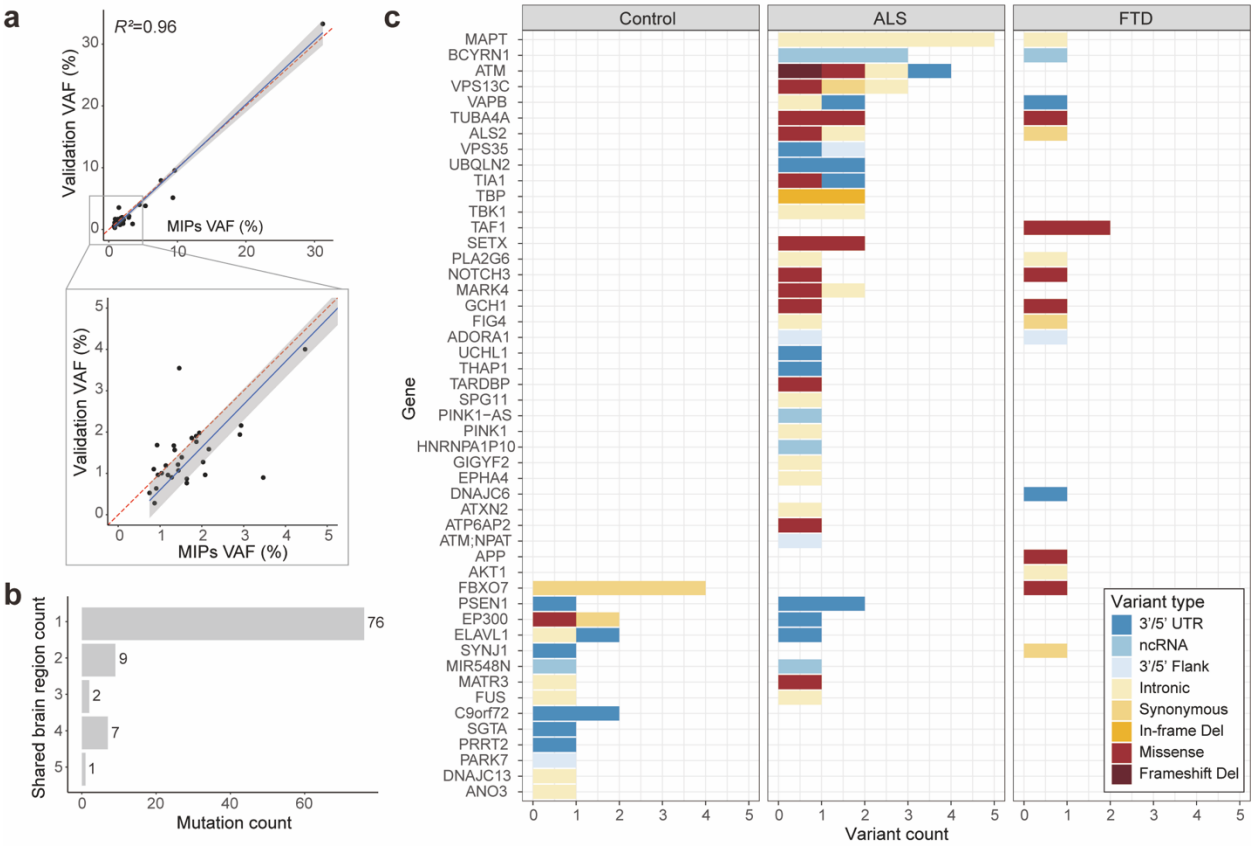
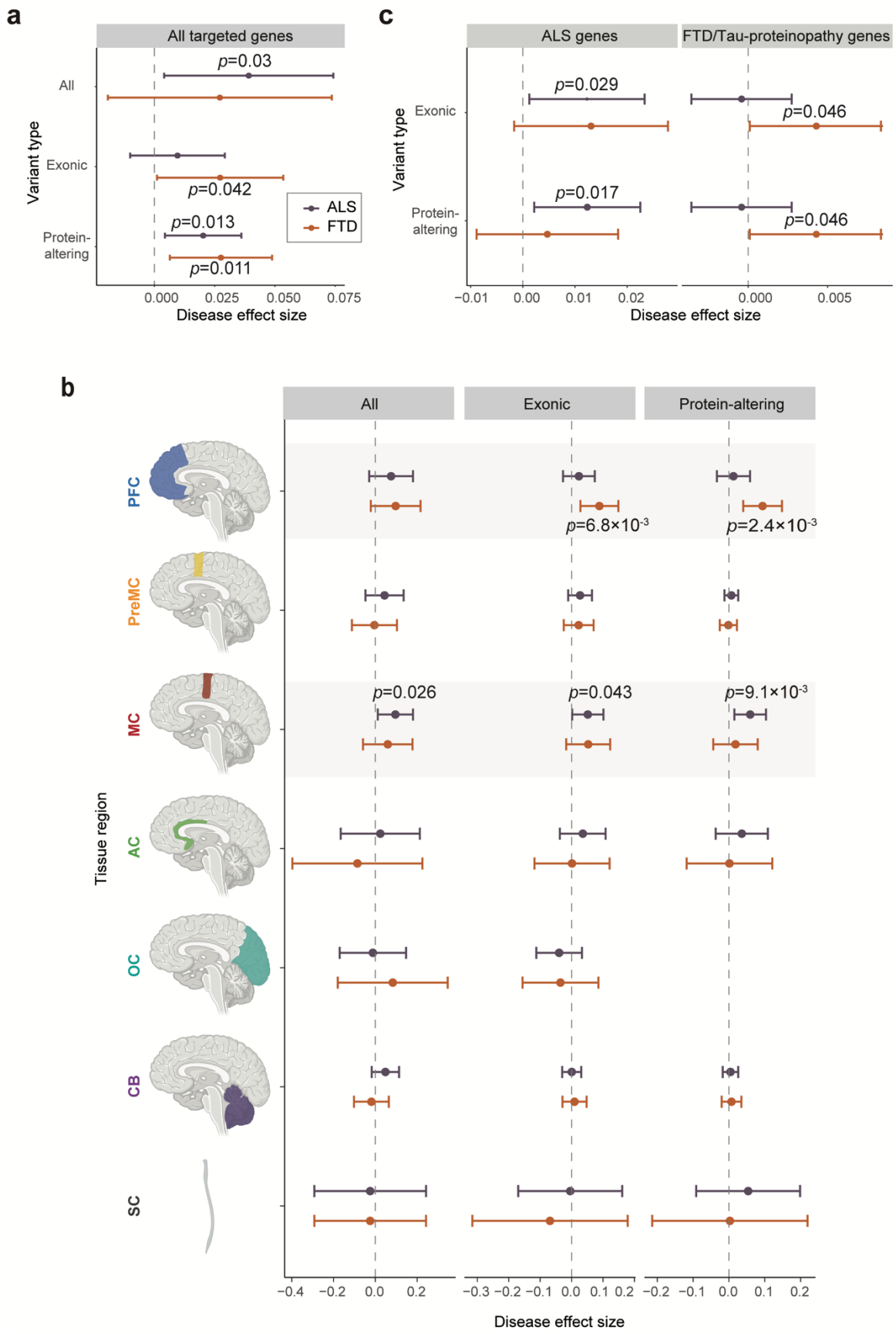
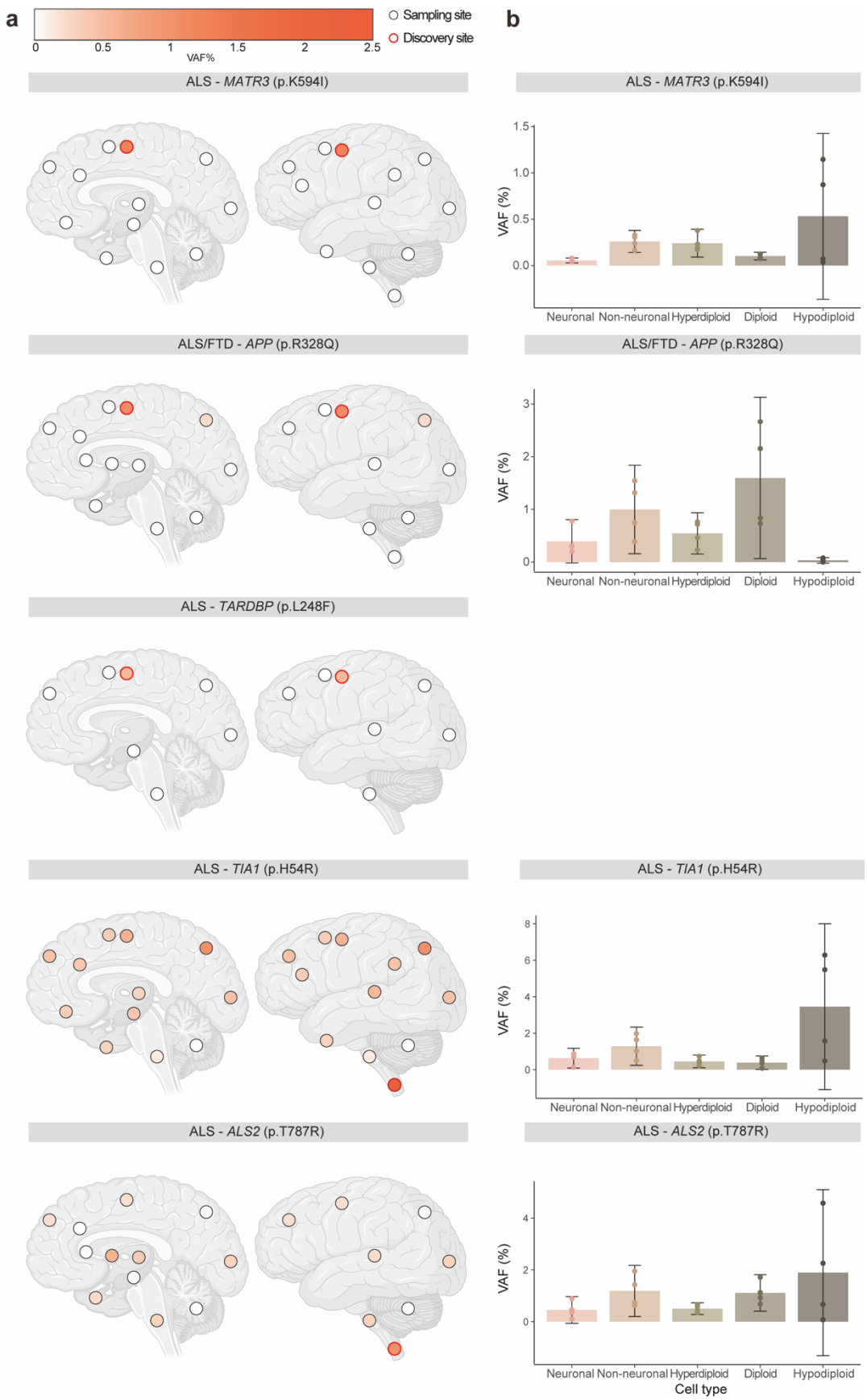


Fig. 3: Somatic variants in MIP sequencing data tend to be focal, protein-altering and are almost exclusively restricted to disease cases. (a) The observed VAFs of somatic variants in amplicon sequencing validation were consistent with the VAFs in original MIP sequencing. Forty somatic variants were validated and included in the plot. (b) Total somatic variant counts classified by the number of brain regions in which a given variant was identified. (c) Distribution of somatic variants in all neurodegenerative genes. Color codes indicate variant types. Note that somatic variants identified in controls are unlikely to alter function, with just one missense mutation (red) and the remaining being synonymous or noncoding substitutions.


a
All targeted genes
p=0.03
All
Exonic
p=0.042
p=0.013
Protein-altering
p=0.011
ALS
FTD
Variant type
0.000
0.025
0.050
0.075
Disease effect size
c
ALS genes
FTD/Tau-proteinopathy genes
p=0.029
Exonic
p=0.046
p=0.017
Protein-altering
p=0.046
Variant type
−0.01
0.00
0.01
0.02
0.000
0.005
Disease effect size
b
All
Exonic
Protein-altering
PFC
PreMC
MC
AC
OC
CB
SC
Tissue region
p=6.8×10-3
p=2.4×10-3
p=0.026
p=0.043
p=9.1×10-3
−0.4
−0.2
0.0
0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
−0.2
−0.1
0.0
0.1
0.2
Disease effect size

448

449 **Fig. 4: Somatic variants are enriched in ALS and FTD cases and disease-related tissue**
450 **regions**. (a) Enrichment of somatic variants in different genomic regions of germline-free ALS
451 and FTD cases compared to normal controls. (b) Enrichment of somatic variants in different
452 brain regions of germline-free ALS and FTD cases compared to normal controls. Significance of
453 enrichment and 95% CI was estimated while controlling for potential confounding factors
454 including average read-depth, sequencing batch, sampled individual using a linear mixed model.
455 (c) Enrichment of exonic and protein-altering somatic variants in two different groups of disease-
456 related genes (ALS genes and FTD/Tau-proteinopathy genes) compared to normal controls
457


a
0
0.5
1
1.5
2
2.5
VAF%
Sampling site
Discovery site
ALS - MATR3 (p.K594I)
ALS/FTD - APP (p.R328Q)
ALS - TARDBP (p.L248F)
ALS - TIA1 (p.H54R)
ALS - ALS2 (p.T787R)
b
ALS - MATR3 (p.K594I)
ALS/FTD - APP (p.R328Q)
ALS - TIA1 (p.H54R)
ALS - ALS2 (p.T787R)
VAF (%)
Neuronal
Non-neuronal
Hyperdiploid
Diploid
Hypodiploid
Cell type

459 **Fig. 5: Pathogenic somatic mutations have restricted regional distributions and are**
460 **enriched in hypodiploid cells**. (a) Regional distribution of VAFs of somatic variants in
461 individual brains and spinal cords. Brain cortex is annotated by Brodmann areas. The color
462 spectrum indicates the VAFs of somatic variants in amplicon sequencing. Dots indicate
463 unavailable regions and white indicates regions without the somatic variants. Red highlight
464 indicates the region of initial detection by MIP sequencing. (b) VAFs of somatic variants in
465 FANS sorted cell types. Five hundred neuronal (NeuN+), non-neuronal (NeuN-), diploid
466 (DAPI), hyperdiploid (High DAPI) and hypodiploid (Low DAPI) cells were each sorted for
467 amplicon sequencing with four replicates. Error bars, 95% CI.
468

a

Individual	Gender	Disease	Onset	Age	Variant	Tissue	VAF
NEUEL133AK6	Female	ALS	Leg	61	*DYNC1H1* (p.R1962C)	Frontal cortex	0.054
						Medial motor cortex	0.036
						Lateral motor cortex	0.042
						Cervical SC	0.021
						Thoracic SC	0.022
						Lumbar SC	0.036
NEUMY028JH4	Female	ALS	Leg	54	*LMNA* (p.H566Y)	Frontal cortex	0.125
						Medial motor cortex	0.179
						Cervical SC	0.172
						Thoracic SC	0.169
						Lumbar SC	0.075

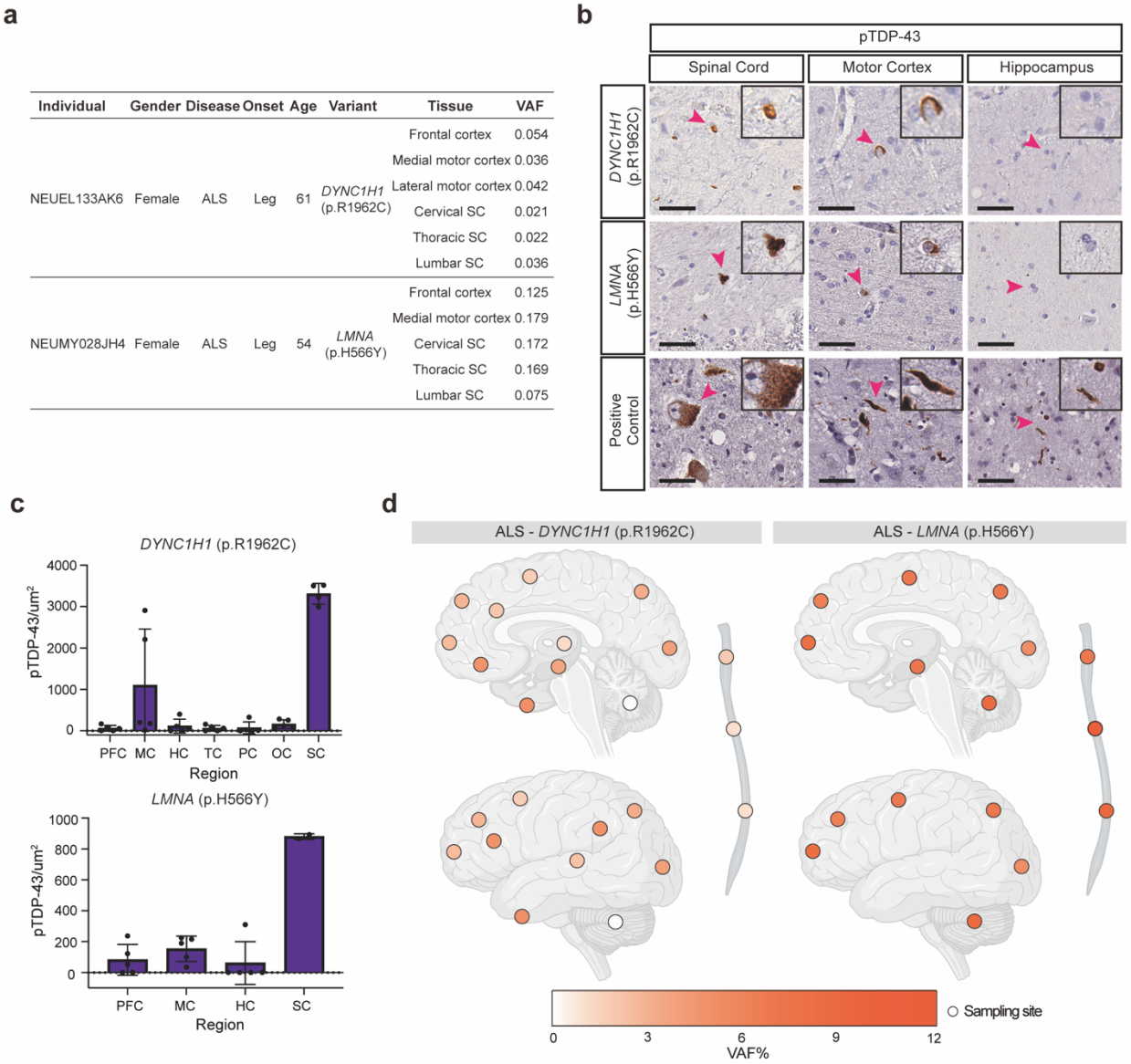


Fig. 6: Somatic variants in *DYNC1H1* and *LMNA* in sALS. (a) Two pathogenic somatic SNVs that were shared by multiple tissue regions of the ALS cases. (b) Sections of the lumbar spinal cord, motor cortex, and hippocampus of the two sALS cases stained with a phospho-TDP43 antibody. Scale bar = 40 um. Arrowheads indicate the cells shown in the insets, which are magnified to twice their original size. (c) Quantification of phospho-TDP43 staining of CNS tissue sections of the two sALS cases with *DYNC1H1* and *LMNA* somatic mutations. Error bars indicate SD (n = 5). PFC: prefrontal cortex. MC: primary motor cortex. HC: hippocampus. TC: temporal cortex. PC: parietal cortex. OC: Occipital cortex. SC: spinal cord. (d) Regional distribution of VAFs of somatic variants in individual brains and spinal cords. Brain cortex is annotated by Brodmann areas. The color spectrum indicates the VAFs of somatic variants in amplicon sequencing. Dots indicate unavailable regions and white indicates regions without the somatic variants.

483
484 **References**

485 1. Ferrari, R., Kapogiannis, D., Huey, E.D. & Momeni, P. FTD and ALS: a tale of two
486 diseases. *Curr Alzheimer Res* **8**, 273-94 (2011).
487 2. Saxon, J.A. *et al.* Examining the language and behavioural profile in FTD and ALS-FTD.
488 *J Neurol Neurosurg Psychiatry* **88**, 675-680 (2017).
489 3. Lagier-Tourenne, C., Polymenidou, M. & Cleveland, D.W. TDP-43 and FUS/TLS:
490 emerging roles in RNA processing and neurodegeneration. *Hum Mol Genet* **19**, R46-64
491 (2010).
492 4. Ling, S.C., Polymenidou, M. & Cleveland, D.W. Converging mechanisms in ALS and
493 FTD: disrupted RNA and protein homeostasis. *Neuron* **79**, 416-38 (2013).
494 5. Ravits, J.M. & La Spada, A.R. ALS motor phenotype heterogeneity, focality, and spread:
495 deconstructing motor neuron degeneration. *Neurology* **73**, 805-11 (2009).
496 6. Kanouchi, T., Ohkubo, T. & Yokota, T. Can regional spreading of amyotrophic lateral
497 sclerosis motor symptoms be explained by prion-like propagation? *J Neurol Neurosurg*
498 *Psychiatry* **83**, 739-45 (2012).
499 7. Eisen, A., Kim, S. & Pant, B. Amyotrophic lateral sclerosis (ALS): a phylogenetic
500 disease of the corticomotoneuron? *Muscle Nerve* **15**, 219-24 (1992).
501 8. Chou, S.M. & Norris, F.H. Amyotrophic lateral sclerosis: lower motor neuron disease
502 spreading to upper motor neurons. *Muscle Nerve* **16**, 864-9 (1993).
503 9. Gromicho, M. *et al.* Spreading in ALS: The relative impact of upper and lower motor
504 neuron involvement. *Ann Clin Transl Neurol* **7**, 1181-1192 (2020).
505 10. Brettschneider, J. *et al.* Stages of pTDP-43 pathology in amyotrophic lateral sclerosis.
506 *Ann Neurol* **74**, 20-38 (2013).
507 11. Brettschneider, J. *et al.* Sequential distribution of pTDP-43 pathology in behavioral
508 variant frontotemporal dementia (bvFTD). *Acta Neuropathol* **127**, 423-439 (2014).
509 12. Polymenidou, M. & Cleveland, D.W. Biological Spectrum of Amyotrophic Lateral
510 Sclerosis Prions. *Cold Spring Harb Perspect Med* **7**(2017).
511 13. Porta, S. *et al.* Patient-derived frontotemporal lobar degeneration brain extracts induce
512 formation and spreading of TDP-43 pathology in vivo. *Nat Commun* **9**, 4220 (2018).
513 14. Laferriere, F. *et al.* TDP-43 extracted from frontotemporal lobar degeneration subject
514 brains displays distinct aggregate assemblies and neurotoxic effects reflecting disease
515 progression rates. *Nat Neurosci* **22**, 65-77 (2019).
516 15. Peng, C., Trojanowski, J.Q. & Lee, V.M. Protein transmission in neurodegenerative
517 disease. *Nat Rev Neurol* **16**, 199-212 (2020).
518 16. De Rossi, P. *et al.* FTLD-TDP assemblies seed neoaggregates with subtype-specific
519 features via a prion-like cascade. *EMBO Rep* **22**, e53877 (2021).
520 17. Tamaki, Y. *et al.* Spinal cord extracts of amyotrophic lateral sclerosis spread TDP-43
521 pathology in cerebral organoids. *PLoS Genet* **19**, e1010606 (2023).
522 18. Kumar, S.T. *et al.* Seeding the aggregation of TDP-43 requires post-fibrillization
523 proteolytic cleavage. *Nat Neurosci* **26**, 983-996 (2023).
524 19. Rosen, D.R. *et al.* Mutations in Cu/Zn superoxide dismutase gene are associated with
525 familial amyotrophic lateral sclerosis. *Nature* **362**, 59-62 (1993).
526 20. Turner, M.R. *et al.* Controversies and priorities in amyotrophic lateral sclerosis. *Lancet*
527 *Neurol* **12**, 310-22 (2013).

528 21. Andersen, P.M. & Al-Chalabi, A. Clinical genetics of amyotrophic lateral sclerosis: what
529 do we really know? *Nat Rev Neurol* **7**, 603-15 (2011).
530 22. Wang, H. *et al.* Smoking and risk of amyotrophic lateral sclerosis: a pooled analysis of 5
531 prospective cohorts. *Arch Neurol* **68**, 207-13 (2011).
532 23. Armon, C. Acquired nucleic acid changes may trigger sporadic amyotrophic lateral
533 sclerosis. *Muscle Nerve* **32**, 373-7 (2005).
534 24. Jamuar, S.S. *et al.* Somatic mutations in cerebral cortical malformations. *N Engl J Med*
535 **371**, 733-43 (2014).
536 25. Proukakis, C. Somatic mutations in neurodegeneration: An update. *Neurobiol Dis* **144**,
537 105021 (2020).
538 26. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion
539 probes. *Nat Biotechnol* **21**, 673-8 (2003).
540 27. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
541 variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
542 28. Renton, A.E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of
543 chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257-68 (2011).
544 29. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding
545 region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245-56
546 (2011).
547 30. Majounie, E. *et al.* Frequency of the C9orf72 hexanucleotide repeat expansion in patients
548 with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study.
549 *Lancet Neurol* **11**, 323-30 (2012).
550 31. Byrne, S. *et al.* Cognitive and clinical characteristics of patients with amyotrophic lateral
551 sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *Lancet*
552 *Neurol* **11**, 232-40 (2012).
553 32. Mahoney, C.J. *et al.* Frontotemporal dementia with the C9ORF72 hexanucleotide repeat
554 expansion: clinical, neuroanatomical and neuropathological features. *Brain* **135**, 736-50
555 (2012).
556 33. van Blitterswijk, M. *et al.* Evidence for an oligogenic basis of amyotrophic lateral
557 sclerosis. *Hum Mol Genet* **21**, 3776-84 (2012).
558 34. Testi, S., Tamburin, S., Zanette, G. & Fabrizi, G.M. Co-occurrence of the C9ORF72
559 expansion and a novel GRN mutation in a family with alternative expression of
560 frontotemporal dementia and amyotrophic lateral sclerosis. *J Alzheimers Dis* **44**, 49-56
561 (2015).
562 35. Baker, M. *et al.* Mutations in progranulin cause tau-negative frontotemporal dementia
563 linked to chromosome 17. *Nature* **442**, 916-9 (2006).
564 36. Cruts, M. *et al.* Null mutations in progranulin cause ubiquitin-positive frontotemporal
565 dementia linked to chromosome 17q21. *Nature* **442**, 920-4 (2006).
566 37. Kuuluvainen, L. *et al.* Oligogenic basis of sporadic ALS: The example of SOD1
567 p.Ala90Val mutation. *Neurol Genet* **5**, e335 (2019).
568 38. Goutman, S.A. *et al.* Emerging insights into the complex genetics and pathophysiology of
569 amyotrophic lateral sclerosis. *Lancet Neurol* **21**, 465-479 (2022).
570 39. Kim, J. *et al.* The use of technical replication for detection of low-level somatic
571 mutations in next-generation sequencing. *Nat Commun* **10**, 1047 (2019).
572 40. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054
573 (2019).

41. Dunn, T. *et al.* Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data. *Bioinformatics* **35**, 1579-1581 (2019).
42. Bizzotto, S. *et al.* Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249-1253 (2021).
43. Lee, J. *et al.* Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res* **46**, W102-W108 (2018).
44. Chung, C. *et al.* Comprehensive multi-omic profiling of somatic mutations in malformations of cortical development. *Nat Genet* **55**, 209-220 (2023).
45. Huang, A.Y. & Lee, E.A. Identification of Somatic Mutations From Bulk and Single-Cell Sequencing Data. *Front Aging* **2**, 800380 (2021).
46. Hadano, S. *et al.* A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat Genet* **29**, 166-73 (2001).
47. Yang, Y. *et al.* The gene encoding alsin, a protein with three guanine-nucleotide exchange factor domains, is mutated in a form of recessive amyotrophic lateral sclerosis. *Nat Genet* **29**, 160-5 (2001).
48. Ferlini, C., Biselli, R., Scambia, G. & Fattorossi, A. Probing chromatin structure in the early phases of apoptosis. *Cell Prolif* **29**, 427-36 (1996).
49. Young, N.A. *et al.* Use of flow cytometry for high-throughput cell population estimates in brain tissue. *Front Neuroanat* **6**, 27 (2012).
50. Hansen, D.V., Hanson, J.E. & Sheng, M. Microglia in Alzheimer's disease. *J Cell Biol* **217**, 459-472 (2018).
51. Rudnik-Schoneborn, S. *et al.* Mutations of the LMNA gene can mimic autosomal dominant proximal spinal muscular atrophy. *Neurogenetics* **8**, 137-42 (2007).
52. Harms, M.B. *et al.* Mutations in the tail domain of DYNC1H1 cause dominant spinal muscular atrophy. *Neurology* **78**, 1714-20 (2012).
53. Tsurusaki, Y. *et al.* A DYNC1H1 mutation causes a dominant spinal muscular atrophy with lower extremity predominance. *Neurogenetics* **13**, 327-32 (2012).
54. Iwahara, N., Hisahara, S., Hayashi, T., Kawamata, J. & Shimohama, S. A novel lamin A/C gene mutation causing spinal muscular atrophy phenotype with cardiac involvement: report of one case. *BMC Neurol* **15**, 13 (2015).
55. Bowerman, M. *et al.* Pathogenic commonalities between spinal muscular atrophy and amyotrophic lateral sclerosis: Converging roads to therapeutic development. *Eur J Med Genet* **61**, 685-698 (2018).
56. Lodato, M.A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98 (2015).
57. Hoang, H.T., Schlager, M.A., Carter, A.P. & Bullock, S.L. DYNC1H1 mutations associated with neurological diseases compromise processivity of dynein-dynactin-cargo adaptor complexes. *Proc Natl Acad Sci U S A* **114**, E1597-E1606 (2017).
58. Poirier, K. *et al.* Mutations in TUBG1, DYNC1H1, KIF5C and KIF2A cause malformations of cortical development and microcephaly. *Nat Genet* **45**, 639-47 (2013).
59. Yang, H. *et al.* De Novo Variants in the DYNC1H1 Gene Associated With Infantile Spasms. *Front Neurol* **12**, 733178 (2021).
60. Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293-8 (2003).
61. Quijano-Roy, S. *et al.* De novo LMNA mutations cause a new form of congenital muscular dystrophy. *Ann Neurol* **64**, 177-86 (2008).

620 62. Boyle, E.A., O'Roak, B.J., Martin, B.K., Kumar, A. & Shendure, J. MIPgen: optimized
621 modeling and design of molecular inversion probes for targeted resequencing.
622 *Bioinformatics* **30**, 2670-2 (2014).
623 63. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
624 *2011* **17**, 3 (2011).
625 64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
626 *arXiv preprint arXiv:1303.3997* (2013).
627 65. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
628 66. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
629 features. *Bioinformatics* **26**, 841-2 (2010).
630 67. Au, C.H., Ho, D.N., Kwong, A., Chan, T.L. & Ma, E.S.K. BAMClipper: removing
631 primers from alignments to minimize false-negative mutations in amplicon next-
632 generation sequencing. *Sci Rep* **7**, 1567 (2017).
633 68. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique
634 Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499
635 (2017).
636 69. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the
637 Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10
638 1-11 10 33 (2013).
639 70. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**,
640 308-11 (2001).
641 71. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**,
642 68-74 (2015).
643 72. Karczewski, K.J. *et al.* The ExAC browser: displaying reference data information from
644 over 60 000 exomes. *Nucleic Acids Res* **45**, D840-D845 (2017).
645 73. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in
646 141,456 humans. *Nature* **581**, 434-443 (2020).
647 74. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-
648 coding variants. *Nature* **493**, 216-20 (2013).
649 75. Scott, E.M. *et al.* Characterization of Greater Middle Eastern genetic variation for
650 enhanced disease gene discovery. *Nat Genet* **48**, 1071-6 (2016).
651 76. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. & Roach, J.C. Kaviar: an accessible
652 system for testing SNV novelty. *Bioinformatics* **27**, 3216-7 (2011).
653 77. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning.
654 *Cell* **176**, 535-548 e24 (2019).
655 78. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting
656 evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
657 79. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*
658 **21**, 577-81 (2003).
659 80. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of
660 Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site
661 SNVs. *Hum Mutat* **37**, 235-41 (2016).
662 81. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous
663 variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-81 (2009).
664 82. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations.
665 *Nat Methods* **7**, 248-9 (2010).

666 83. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes.
667 *Genome Res* **19**, 1553-61 (2009).
668 84. Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates
669 disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-6 (2010).
670 85. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations:
671 application to cancer genomics. *Nucleic Acids Res* **39**, e118 (2011).
672 86. Shihab, H.A. *et al.* Predicting the functional, molecular, and phenotypic consequences of
673 amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65 (2013).
674 87. Shihab, H.A. *et al.* An integrative approach to predicting the functional effects of non-
675 coding and coding sequence variation. *Bioinformatics* **31**, 1536-43 (2015).
676 88. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional
677 effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
678 89. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for
679 nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125-37
680 (2015).
681 90. Zook, J.M. *et al.* Integrating human sequence data sets provides a resource of benchmark
682 SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-51 (2014).
683 91. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
684 (2013).
685 92. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-
686 generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
687 93. Huang, A.Y. *et al.* MosaicHunter: accurate detection of postzygotic single-nucleotide
688 mosaicism through next-generation sequencing of unpaired, trio, and paired samples.
689 *Nucleic Acids Res* **45**, e76 (2017).
690 94. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human
691 genomes. *Nature* **491**, 56-65 (2012).
692 95. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep
693 sequencing of human exomes. *Science* **337**, 64-9 (2012).
694 96. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*
695 **536**, 285-91 (2016).
696 97. Evrony, G.D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and
697 somatic mutation in the human brain. *Cell* **151**, 483-96 (2012).
698 98. Nolan, M. *et al.* Quantitative patterns of motor cortex proteinopathy across ALS
699 genotypes. *Acta Neuropathol Commun* **8**, 98 (2020).
700

701

Methods

Tissue sources and sample preparation

Fresh frozen postmortem human brain and spinal cord tissues were collected by the Massachusetts Alzheimer's Disease Research Center, Oxford Brain Bank, Target ALS Foundation, and NIH NeuroBioBank (Supplementary Table 1) 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. Sporadic ALS and FTD cases were selected based on available clinical records. ALS and FTD cases without clear recording of family histories were also selected if the age of death was above 45 years old. gDNA of these tissue samples was extracted using the EZ1 Advanced XL (Qiagen) system followed by an additional purification using AMPure XP beads (Beckman Coulter).

MIP panel design

A double-stranded DNA MIP panel targeting 1.4Mb across exons and exon-intron junctions of 88 neurodegenerative genes was designed using custom scripts incorporating MIPgen[62] using the human reference genome, hg19, with Mly1 restriction sites masked with 'N' using bedtools. The final panel of 26,439 MIPs captures an average fragment length of 209bp, including the extension and ligation arms to ensure overlapping of the forward and reverse sequencing read. The panel successfully targets 92.7% of bases including flanking intronic regions, with >98% of exonic bases covered with an average of at least 2 unique MIPs. All MIPs were designed to include a custom backbone consisting of primer binding sites and dual 5nt unique molecular indexes (UMI). MIPs were rebalanced in the pool based on the percent of GC content within the regions. Common primer binding and Mly1 restriction enzyme sites were added to both ends of the oligo sequences to enable blunt-end removal of the primer binding sites. The forward and reverse compliment sequences were printed into a single ssDNA pool by CustomArray (Bothell, WA). The resulting panel was amplified at a low cycle number (12X), digested with Mly1 enzyme for 12 hrs at 37C, and purified using Qiagen Nucleotide removal kit.

MIP capture and library construction

Two hundred fifty ng of gDNA was first hybridized in a 15 ul reaction with 1.5 ul of Ampligase® 10X Reaction Buffer (VWR), 1.5 ul of the reverse blocking oligo (5'-NNNNGAAGTCGAAGGGCTATAGGCTGCCATCACANNNN-3') and the MIP pool at 63 nM for
10 min at 95 °C and 24 hrs at 60 °C. Gap-fill/ligation was then performed by adding 1 unit of Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher), 4 units of Ampligase® DNA Ligase (Epicentre), 0.2 ul of Ampligase® 10X Reaction Buffer, 0.6 ul of dNTPs (10 mM) and 1 ul of nuclease-free water to the MIP capture product and incubated at 60 °C for 1 hr. For exonuclease digestion, 50 units of Exonuclease III (Thermo Fisher), 10 units of Exonuclease I (Thermo Fisher), 0.2 ul of Ampligase® 10X Reaction Buffer (VWR), and 2.05 ul of nuclease-free water was added to the Gap-fill/ligation product, which was incubated for 40 min at 37 °C and 5 min at 95 °C. Ten ul of the captured library is amplified in a 50 ul final reaction by adding 1 unit of Phusion Hot Start II DNA Polymerase (Thermo Fisher), 10 ul of 5X HF buffer, 1 ul of dNTPs (10mM), 1 ul of the universal MIP barcode forward primer (10 uM), 1 ul of the individual barcode reserve primer (10 uM) and 26.5 ul of nuclease-free water. MIP library

748 amplification was then performed under the following conditions: 98 °C for 30 s; 16 cycles of
749 98 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s; 72 °C for 2 min. MIP library was then purified
750 using 2X AMPure XP Beads (Beckman Coulter,) and quantified by the Quant-iT™ dsDNA
751 Assay HS Kit (Thermo Fisher). Ninety-six MIP libraries were pooled together and sequenced on
752 one lane of Illumina Hiseq X.
753
754 **Pre-processing and read mapping of MIP sequencing data**
755 MIP sequencing primers were removed first from the raw FASTQ files using Cutadapt[63] (v2.4,
756 5' adapter of the first read: CATACGAGATCCGTAATCGGGAAGCTGAAG, 3' adapter of the
757 first read: ACACTACCGTCGGATCGTGCGTGT, 5' adapter of the second read:
758 GCTAAGGGCCTAACTGGCCGCTTCACTG, 3' adapter of the second read:
759 CTTCAGCTTCCCGATTACGGATCTCGTATG). Trimmed reads were aligned to the human
760 reference genome (GRCh37) using BWA-mem[64] (v0.7.15) and sorting and indexing were
761 performed using samtools[65] (v1.3.1). From the aligned BAM file, off-target reads were removed
762 by checking the overlaps with the target regions using bedtools[66] (v2.26.0). MIP arm regions
763 were masked by soft-clipping for each read using BAMClipper[67] (v1.0.1). Unique molecular
764 identifier (UMI) information was extracted, and then mapped reads were deduplicated based on
765 the mapping coordinate and the shared UMI using UMI-tools[68] (v1.0.0). Base quality score
766 recalibration and local realignment were performed using the Genome Analysis Toolkit (GATK,
767 v3.7)[69], generating final analysis-ready BAMs.
768
769 **Variant calling for pathogenic germline mutations**
770 Initial candidates of germline SNVs and indels were identified using GATK HaplotypeCaller
771 with default parameter settings. Low-quality candidates were filtered out if any of the following
772 conditions is not satisfied: 1) ≥ 10 variant-supporting reads, 2) ≥ 20 total read-depth at the
773 variant site, 3) VAF ≥ 0.3, 4) GATK QUAL ≥ 50, and 5) identified in all brain regions from the
774 same individual except for the samples failed to cover the variant site (<10 reads). Possible
775 pathogenic germline variants were further selected by satisfying all the following conditions: 1)
776 present in less than 0.1% of the population in any ethnic group of public databases including
777 dbSNP[70], the 1000 Genomes Project[71], the Exome Aggregation Consortium (ExAC)[72], the
778 Genome Aggregation Database (gnomAD)[73], the NHLBI Exome Sequencing Project
779 (ESP6500)[74], the Greater Middle East variome project (GME)[75], and Kaviar database[76], 2)
780 candidates observed only in disease or control groups but not in both, 3) possible protein-altering
781 candidates (missense, nonsense, frame-shift, or splicing variants), and 4) affecting 30 ALS- and
782 FTD-related genes. Pathogenicity prediction module (see computational prediction of variant
783 pathogenicity section below) was then applied to the remaining candidates, and predicted
784 pathogenic variants were reported as final pathogenic germline mutations. ANNOVAR[27] was
785 used to annotate the genomic region, affected genes, population allele frequency, and exonic
786 variant functions. SpliceAI[77] was additionally utilized to identify more splice-altering variants.
787 Candidates with delta score > 0.5 were considered to be possible splicing variants.
788
789 ***C9ORF72* repeat expansion genotyping**
790 Repeat-primed PCR (RP-PCR) of the *C9ORF72* repeat expansion was performed in a 30 ul PCR
791 reaction with 150 ng of gDNA, 15 ul of 2X FastStart™ PCR Master (Roche), 2 ul of DMSO, 5
792 ul of 5X Q-solution (Qiagen), 1 ul of 5 mM 7-deaza-dGTP (NEB), 1 ul of 25 mM $MgCl_2$
793 (Qiagen) and 1 ul of the primer mix (40 uM of the Forward primer: 5'-/56-

794 FAM/AGTCGCTAGAGGCGAAAGC-3', 20 uM of the Reverse primer: 5'-
795 TACGCATCCCAGTTTGAGACGGGGGCCGGGGCCGGGGCCGGGG-3' and 40 uM of the
796 Anchor/tail primer: 5'-TACGCATCCCAGTTTGAGACG-3'. The reaction was performed with
797 touchdown PCR cycling conditions consisting of 15 minutes at 95°C, followed by cycles of 94°C
798 for 1 minute, annealing starting at 70°C for 1 minute, and extension at 72°C for 3 minutes,
799 ending with a final extension step of 10 minutes at 72°C. The annealing temperature was
800 decreased in 2°C steps as follows: 70°C for two cycles, 68°C for three cycles, 66°C for four
801 cycles, 64°C for five cycles, 62°C for six cycles, 60°C for seven cycles, 58°C for eight cycles,
802 and 56°C for five cycles. The RP-PCR products were separated by the SeqStudio Genetic
803 Analyzer (Thermo Fisher) with the GeneScan™ 600 LIZ™ Dye Size Standard (Thermo Fisher).
804 Results of fragment sizes were analyzed by Peak Scanner™ Software v1.0 (Thermo Fisher).
805
806 **Somatic variant calling from MIP sequencing data**
807 Three different callers RePlow (v1.1.0)[39], Mutect2 (v4.1.5)[40], and Pisces (v5.2.11)[41] were used to
808 generate initial candidate sets. Each sample was analyzed by all three callers with the single-
809 sample mode. Default parameter settings were used except for the adjustments for disabling the
810 coverage limit. Variants that passed all the filters from each caller were used to make three
811 different initial sets. Candidates identified by only one caller were discarded, and those called at
812 least two callers were retained as a double-call set. For indels, double-calls between Mutect2 and
813 Pisces were used as somatic indel candidates since RePlow does not support indel detection. For
814 SNVs, among double-calls Mutect2-Pisces pairs were additionally filtered out due to high false
815 positive rates and low validation rates in the benchmarking data set (Supplementary Fig. 3).
816 Remaining RePlow-based SNV double-calls and indel candidates were subject to multi-step
817 variant filters to further remove false positive candidates.
818 Unlike germline variant calling, somatic variant calling aims to reliably detect low-VAF
819 mutations up to ~0.5%, which requires enough supporting evidence to control the false positive
820 rate. Calling thresholds such as variant-supporting read count, read-depth at the variant site, and
821 average base-call quality were determined based on the benchmarking data. Somatic variants
822 were selected satisfying all the following conditions: 1) ≥ 50 total read-depth at the variant site,
823 2) ≥ 15 variant-supporting reads excluding the reads with the variant allele on their probe-arm
824 regions, 3) > 30 average base-call quality of variant allele, 4) ≥ 2 different types of variant-
825 supporting amplicons, 5) 0.001 ≤ VAF ≤ 0.4, 6) ≤ 3 variant candidates within 20 bp window
826 from the same sample, 7) present in less than 0.1% of the population in any ethnic group of
827 public databases and 8) observed in < 5 different individuals.
828 We additionally found that low-level contamination of DNA from another sample occurred in a
829 few samples. Germline variants from the contaminant mimicked low-VAF somatic mutations
830 and generated false positive calls. We therefore implemented a module to identify low-level
831 contamination and filter out candidates that originated from the contaminant. By comparing a
832 somatic candidate set from a given sample with the germline call set of every individual, sample
833 contamination was determined if the given sample has ≥ 40 low-VAF somatic candidates that are
834 also observed in a specific individual as germline variants. In this case, germline variants of the
835 matched individual are considered to be possible sources of false positive calls and all somatic
836 candidates that are matched with these germline variants from the individual were filtered out.
837 The remaining candidates were reported as final somatic variants.
838 Pathogenic somatic variants were further annotated with similar criteria for selecting pathogenic
839 germline variants. Among final candidates, variants that are 1) observed only in disease or

control groups but not in both, 2) possible protein-altering variants, and 3) affecting ALS- and FTD-related genes were selected and applied for the pathogenicity prediction module. ANNOVAR and SpliceAI were utilized to annotate variants with various genomic information and detect additional splice-altering variants, respectively.

Computational prediction of variant pathogenicity

Pathogenicity prediction module was applied to filtered germline and somatic variants to refine the pathogenic candidate sets. Variants that were previously reported as benign/likely benign in the clinical databases (ClinVar[78] and Human Gene Mutation Database[79]) were excluded from the pathogenic candidate set. Nonsense, frameshift, and canonical splicing variants (±1-2 splice sites) were assumed to disrupt gene function and were included in the pathogenic set. For missense variants, the dbNSFP database[80] was utilized to adopt multiple computational algorithms (SIFT[81], PolyPhen2[82], LRT[83], MutationTaster[84], MutationAssessor[85], FATHMM[86], FATHMM-MKL[87], PROVEAN[88], MetaSVM[89], MetaLR[89]), considering damaging effects at different levels such as biochemical property, protein structure, and evolutionary conservation. Categorical prediction results of each algorithm were delivered by ANNOVAR. A missense variant was selected to be pathogenic if at least three different algorithms predicted damaging effects (deleterious for SIFT, LRT, FATHMM, PROVEAN, MetaSVM and MetaLR; probably damaging for PolyPhen2; disease_causing for MutationTaster), while excluding possibly/likely damaging predictions from the counts for more conservative selection. For ALS/FTD-related genes, previously reported inheritance patterns (dominant/recessive) were carefully checked. For recessive genes, two independent mutations in the same gene were required to determine whether a given individual was affected by pathogenic mutations.

Benchmarking with spike-in datasets

Two Coriell cell lines (GM12878 and GM24695) were used to generate a spike-in data. Extracted DNA were mixed at five different levels to mimic low-level somatic mutations, targeting the VAFs of 0.5%, 1%, 2.5%, 5%, and 10%. Genomic DNA from GM12878 cells was spiked into DNA from GM24695, therefore unique germline SNPs in GM12878 were served as somatic mutations. Genomic position and genotype information for germline SNPs of Coriell samples were obtained from NIST high-confidence call sets[90]. A total of 165 SNPs (57 homozygous and 108 heterozygous SNPs) covered by our designed MIP panel were used as the benchmark variant set. RePlow, Mutect2, Pisces, and their combinations were tested. Detected mutations not in the benchmark set were considered to be false positives, except for GM24695 germline SNPs.

Somatic variant calling from RNA-seq data

Raw bam files of RNA-seq and matched WGS data for sALS and control cases of the New York Genome Center ALS Consortium were obtained from the New York Genome Center. RNA-seq reads extracted from raw bam files were aligned to the GRCh38 human reference genome by STAR (v2.5.0a)[91] in the two-pass mode with the reference gene annotation (Gencode version 39). The aligned bam files were processed by Picard (v1.138) to remove duplicates, and then by GATK (v3.6)[92] for SplitNCigarReads, indel realignment, and base quality recalibration. We further excluded reads that were improperly paired or with ambiguous alignment. Somatic SNVs were called by RNA-MosaicHunter (v1.0) with default parameters (https://gitlab.aleelab.net/august/rna-mosaichunter; manuscript in submission). Derived from

886 MosaicHunter[93], which was designed for somatic mutation calling in DNA sequencing, RNA-
887 MosaicHunter incorporates a Bayesian genotyper and a series of empirical filters to
888 systematically distinguish somatic mutations from technical artifacts and germline mutations,
889 with 59% sensitivity and 94% precision benchmarked using cancer datasets. Specifically,
890 germline mutations identified from the matched WGS data from the same individual were
891 excluded. We excluded A-to-G candidates because they are most likely led by the widespread A-
892 to-I(G) RNA editing events in the human genome. To remove recurrent artifacts, we only
893 considered exonic candidates that were called in one or two individuals. We further excluded
894 candidates present in human polymorphism databases including dbSNP[70], the 1000 Genomes
895 Project[94], the Exome Sequencing Project[95], and the Exome Aggregation Consortium[96].
896
897 **Nuclei isolation and whole genome amplification**
898 Isolation of total (DAPI+), neuronal (NeuN+), non-neuronal (NeuN-), and damaged (low DAPI)
899 nuclei were achieved by FANS together with nuclear staining of NeuN (Millipore, MAB377)
900 and DAPI following a previously published study[97]. Five hundred nuclei of each cell population
901 were sorted into wells of 96-well plates.
902 Sorted nuclei were subjected to genome amplification using the Primary Template-directed
903 Amplification kit (BioSkryb, 100136) following the manufacturer's protocol.
904
905 **Amplicon sequencing**
906 Primer sets targeting each identified somatic SNV were designed using BatchPrimer3
907 (Supplementary Table 11). Amplicon was amplified for 25 cycles in a 50 ul PCR reaction with
908 50 ng of gDNA, 1 unit of Phusion Hot Start II DNA Polymerase (Thermo Fisher), 10 ul of 5X
909 HF buffer, 1 ul of dNTPs (10mM) and 10 ul of each primer (10 uM). Amplicon PCR products
910 were then purified by a 0.65X + 1.05X double size selection with AMPure XP Beads (Beckman
911 Coulter, A63882). Purified amplicons were then pooled based on the concentrations measured by
912 the Quant-iT™ dsDNA Assay HS Kit (Thermo Fisher) and sequenced using Amplicon-EZ
913 (Genewiz).
914
915 **Burden analysis of somatic mutations using linear mixed model**
916 Linear mixed-effect regression model was used to compare somatic mutation burden between
917 clinical conditions while accounting for other covariates that may affect the burden. Clinical
918 conditions and covariates of interest (e.g. age, gender, sequencing depth) were modeled as fixed
919 effects and the batch and individual (donor) information were modeled as random effects,
920 considering the uncertainty caused by sample clusters from the same origin (donor or batch).
921 Somatic mutation count in each sample was normalized per megabase pair and modeled as a
922 dependent variable. A covariate with a p-value < 0.05 was considered to be significant, based on
923 a t-test using the Satterthwaite approximation of degrees of freedom. To test the burden of
924 somatic mutations in different genomic regions, a linear mixed model was fitted to the mutation
925 counts of specific type (e.g. exonic). To test the burden of somatic mutations in different brain
926 regions, samples were first divided by the sequenced region and then a linear mixed model was
927 fitted for each region group.
928
929 **Immunohistochemistry**
930 Immunohistochemistry was performed using DAB (3,3'-Diaminobenzidine) detection as
931 previously described[98]. Briefly, 7μm formalin-fixed, paraffin-embedded (FFPE) sections were

932 dewaxed using citrisolve, before being rehydrated through decreasing concentrations of ethanol.
933 Antigen retrieval was performed using sodium citrate buffer pH 6.0 at 121°C for 15 mins.
934 Endogenous peroxidases were blocked using 3% hydrogen peroxide solution, and non-specific
935 binding was blocked using 10% normal goat serum. Sections were then incubated overnight at
936 4°C with primary antibody (pTDP-43 mouse monoclonal, CosmoBio CAC-TIP-PTD-P03,
937 1:10,000). After washing with TBS-Triton, sections were incubated with a Horseradish
938 peroxidase (HRP)-conjugated Goat anti-mouse secondary (Dako) for one hour at room
939 temperature. HRP signal was detected using DAB substrate (Dako) applied for 15 minutes.
940 Counterstaining was performed using Coles hematoxylin for 1 minute. Sections were then
941 dehydrated, cleared using citrisolve, and mounted using glass coverslips. All sections were
942 viewed using a Leica upright light microscope and assessed for section quality prior to whole-
943 slide digital scanning.
944
945 **Quantification of p-TDP43 burden by immunohistochemistry**
946 Stained sections were scanned using a NanoZoomer whole-slide digital imager at 40X
947 magnification. Images were then visualized and quantified using QuPath image analysis software
948 and algorithms described previously[98]. Briefly, for cortical/cerebellar sections 5 ROI measuring
949 3mm2 (1000 x 3000µm) were placed equidistantly around a single gyrus with the short end of
950 the ROI placed at the pial surface. Pathology was then quantified using a positive pixel count
951 within each ROI and measurements were averaged to provide an output of positive pixels/mm2.
952 For spinal cord sections, square ROI (2.25mm^2) was placed on each side of the central canal
953 within the anterior horn and measurements were averaged.
954
955 **Data availability**
956 The bulk RNA-seq data for the NYGC ALS Consortium samples can be obtained upon request
957 through the NYGC. The MIP targeted gene panel sequencing data generated in this study will be
958 deposited to dbGaP with controlled use conditions set by human privacy regulations. Germline
959 and somatic mutations identified and validated in this study are listed in the supplementary
960 tables.
961
962 **Code availability**
963 The source code and default configuration file of RNA-MosaicHunter are available at
964 https://gitlab.aleelab.net/august/rna-mosaichunter.git. The implemented codes for preprocessing
965 of MIP sequencing data, statistical test, and visualization will be available before publication.
966
967 **Acknowledgements**
968 We thank the Massachusetts Alzheimer's Disease Research Center, Oxford Brain Bank, Target
969 ALS Foundation (Biobank Core Facility at St. Joseph's Hospital and Barrow Neurological
970 Institute, Georgetown Brain Bank, Eleanor and Lou Gehrig ALS Center at Columbia University
971 and UCSD ALS bank) and NIH NeuroBioBank (Harvard Brain Tissue Resource Center, Mount
972 Sinai/JJ Peters VA Medical Center NIH Brain and Tissue Repository, Brain Endowment Bank of
973 University of Miami, University of Pittsburgh Neuropathology Brain Bank, University of
974 Maryland Brian and Tissue Bank and UCLA Human Brain and Spinal Fluid Resource Center)
975 for providing fresh frozen human tissues. We thank the Target ALS Human Postmortem Tissue
976 Core, New York Genome Center for Genomics of Neurodegenerative Disease, Amyotrophic
977 Lateral Sclerosis Association and TOW Foundation for providing the bulk RNA-seq data. We
978 thank the donors and families for their contributions, and J. E. Neil and J. Gonzalez for

assistance with tissue procurement. We thank the Research Computing group at Harvard Medical School and Boston Children's Hospital. The brains in Fig. 5 and Fig. 6 were illustrated by A. Lai with input from the authors. This work was supported by the PRMRP Discovery Award W81XWH2010028 (Z.Z.); the Edward R. and Anne G. Lefler Center postdoctoral fellowship (Z.Z.); the American Heart Association Career Development Award 23CDA1046074 (Z.Z.); the National Research Foundation of Korea (NRF) 2022R1C1C1010430 (J.K.); the Alzheimer's Association research fellowship (A.Y.H.); R56 AG079857 (A.Y.H., C.A.W. and E.A.L.); A Cullen Education and Research Foundation Young Investigator Award from the Healey Center (M.N.); a Holloway Postdoctoral Fellowship from the Association for Frontotemporal Degeneration (M.N.); K08 AG065502 (M.B.M.); donors of the Alzheimer's Disease Research program of the BrightFocus Foundation A20201292F (M.B.M.); the Doris Duke Charitable Foundation Clinical Scientist Development Award 2021183 (M.B.M.); K01 AG051791 (E.A.L.); the Suh Kyungbae Foundation (E.A.L.), DP2 AG072437 (E.A.L.); R01 NS032457 (C.A.W.); R01 AG070921 (C.A.W. and E.A.L.); a Massachusetts Alzheimer's Disease Research Center pilot grant (C.L.-T. and C.A.W.); and the Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation (C.A.W. and E.A.L.). C.L.-T. is supported by the Araminta Broch-Healey Endowed Chair in ALS. C.A.W. is an Investigator of the Howard Hughes Medical Institute. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Z.Z., J.K. and A.Y.H. conceived and designed the study. Z.Z. performed tissue processing, MIP panel sequencing, cell sorting and amplicon sequencing. J.K. performed bioinformatic analysis for MIP sequencing data and validation with assistance from R.D. and T.S.. A.Y.H. performed bioinformatic analysis for bulk RNA-seq data with assistance from J.P.. M.N. optimized and performed immunofluorescent imaging and quantification, and generated data shown in this manuscript. Z.Z., M.M. and R.D. designed the MIP panel. B.C., K.M., and R.Y. helped with tissue processing and amplicon sequencing. C.K. provided technical support for MIP sequencing. J.E.N. contributed tissue procurement and ethics expertise. T.O. and J.R. provided immunofluorescent images and interpretation of disease pathology. L.W.O and O.A. provided fresh frozen human tissues and interpretation of disease pathology. C.A.W., E.A.L. and C.L.-T. supervised the study. Z.Z., J.K., A.Y.H., C.A.W., E.A.L. and C.L.-T. wrote the manuscript.