**Article**
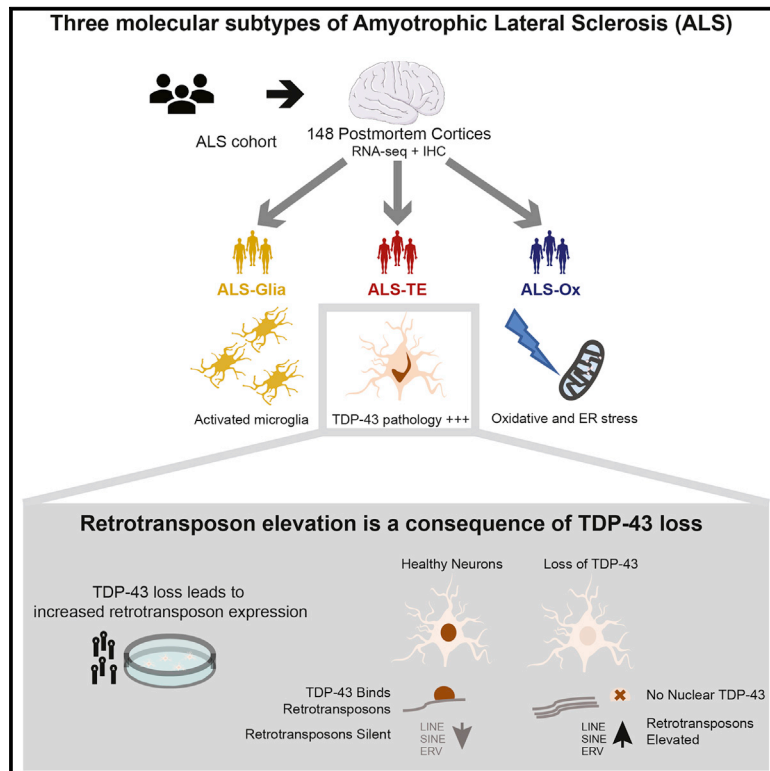
# Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia

## Graphical Abstract



## Authors

Oliver H. Tam, Nikolay V. Rozhkov, Regina Shaw, ..., John Ravits, Josh Dubnau, Molly Gale Hammell

## Correspondence

mhammell@cshl.edu

## In Brief

Tam et al. present transcriptome profiling results from a large set of amyotrophic lateral sclerosis (ALS) patient cortex samples, finding 3 distinct groups. Two ALS subtypes are marked by gene pathways previously associated with ALS disease, while a third group shows elevated retrotransposon expression linked to TDP-43 pathology.

## Highlights

- ALS patient cortex samples can be grouped into 3 distinct expression profiles

- Aberrant pathways include stress, inflammation, and TDP-43 pathology

- Elevated retrotransposon expression is correlated with TDP-43 pathology

- TDP-43 directly binds retrotransposons in cells and contributes to their silencing

**CellPress**

# Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia

Oliver H. Tam,[1,9] Nikolay V. Rozhkov,[1,9] Regina Shaw,[1,9] Duyang Kim,[2] Isabel Hubbard,[2,10] Samantha Fennessey,[2] Nadia Propp,[2] The NYGC ALS Consortium, Delphine Fagegaltier,[2] Brent T. Harris,[3,4] Lyle W. Ostrow,[3,5] Hemali Phatnani,[2,3] John Ravits,[3,6] Josh Dubnau,[3,7,8] and Molly Gale Hammell[1,3,11,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2]Center for Genomics of Neurodegenerative Disease, New York Genome Center, New York, NY 10013, USA
[3]The NYGC ALS Consortium
[4]Department of Neurology, Georgetown University Medical Center, Washington, DC 20007, USA
[5]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[6]Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA
[7]Department of Anesthesiology, Stony Brook University, Stony Brook, NY 11794, USA
[8]Department of Neurobiology and Behavior, Stony Brook University, Stony Brook, NY 11794, USA
[9]These authors contributed equally
[10]Present address: Department of Neurobiology and Behavior, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY 11794, USA
[11]Lead Contact
*Correspondence: mhammell@cshl.edu
https://doi.org/10.1016/j.celrep.2019.09.066

## SUMMARY

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized by the progressive loss of motor neurons. While several pathogenic mutations have been identified, the vast majority of ALS cases have no family history of disease. Thus, for most ALS cases, the disease may be a product of multiple pathways contributing to varying degrees in each patient. Using machine learning algorithms, we stratify the transcriptomes of 148 ALS postmortem cortex samples into three distinct molecular subtypes. The largest cluster, identified in 61% of patient samples, displays hallmarks of oxidative and proteotoxic stress. Another 19% of the samples shows predominant signatures of glial activation. Finally, a third group (20%) exhibits high levels of retrotransposon expression and signatures of TARDBP/TDP-43 dysfunction. We further demonstrate that TDP-43 (1) directly binds a subset of retrotransposon transcripts and contributes to their silencing in vitro, and (2) pathological TDP-43 aggregation correlates with retrotransposon de-silencing in vivo.

## INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a fatal progressive neurodegenerative disorder with no known cure, and only two Food and Drug Administration (FDA)-approved treatments that appear to mildly slow disease progression. ALS is a largely sporadic disease, with 90% of patients carrying no known genetic mutation

or family history of disease. Large-scale patient sequencing studies have identified a growing number of genes in which mutations are linked to ALS (Chia et al., 2018; Nicolas et al., 2018; van Rheenen et al., 2016). The most common ALS-associated mutations are repeat expansions in the intronic region of C9orf72, while mutations in well-known ALS-associated genes such as SOD1 and TARDBP are typically present in fewer than 2% of all ALS patients (Chia et al., 2018).

Mutations in the TARDBP gene (that generates the TDP-43 protein) are rare in ALS, yet nearly all ALS patients exhibit cytoplasmic aggregates of TDP-43 in the affected tissues (Arai et al., 2006; Neumann et al., 2006). TDP-43 has known roles in RNA splicing, stability, and small RNA biogenesis (Cohen et al., 2011). Recently, several studies have suggested that TDP-43 also plays a role in regulating the activity of retrotransposons (Chang and Dubnau, 2019; Krug et al., 2017; Li et al., 2015; Saldi et al., 2014). Retrotransposons, a subset of transposable elements (TEs), are genomic parasites capable of inserting new copies of themselves throughout the genome by a process called retrotransposition. Previous work from our lab and others has shown that TDP-43 represses retrotransposon transcripts at the RNA level in animal models of TDP-43 pathology (Krug et al., 2017; Li et al., 2012). However, a role for TDP-43 in general retrotransposon silencing has not been demonstrated, nor whether TDP-43 pathology in ALS patients correlates with retrotransposon de-silencing. Of note, prior studies have identified a link between retrotransposon expression and repeat expansion in another ALS-linked gene, C9orf72 (Prudencio et al., 2017). Finally, contrasting studies either failed to find an enrichment for elevated levels of the endogenous retrovirus HERV-K in a smaller sample of ALS tissues (Mayer et al., 2018) or suggested that TDP-43 may activate HERV-K transcription rather than silencing this particular retrotransposon (Li et al., 2015). These

studies left open the question of whether retrotransposon silencing is a conserved role for TDP-43 and whether retrotransposon de-silencing would be expected in human tissues with TDP-43 dysfunction.

Here, we show that robust retrotransposon de-silencing occurs in a distinct subset of ALS patient samples, and this is associated with TDP-43 dysfunction. Unbiased machine learning algorithms identified three distinct ALS patient molecular subtypes within the large ongoing sequencing survey by the NYGC ALS Consortium. These subtypes represented both ALS disease-implicated signatures as well as additional correlated pathways. The largest subgroup of patients (61%) showed evidence of oxidative and proteotoxic stress. A second subgroup (19%) displayed strong signatures of glial activation and inflammation. A third subgroup (20%) was marked by retrotransposon re-activation as a dominant feature. We further validated the correlation between TDP-43 pathology and retrotransposon de-silencing in a second independent cohort of postmortem tissue samples, which also recapitulated the three distinct molecular subtypes. These subtypes may reflect different predominant aberrant cellular mechanisms contributing to ALS pathogenesis, and thus suggest specific therapeutic strategies may have greater relevance to distinct sets of sporadic ALS patients.

## RESULTS

### Evidence for Distinct Molecular Subtypes in ALS Patient Samples

The NYGC ALS Consortium has gathered deeply sequenced transcriptomes from the frontal cortex of 77 ALS patients as well as 18 neurological and non-neurological controls (Figure 1A). For some patients, multiple samples were taken from various regions of the frontal cortex, including motor cortex, such that 148 total transcriptomes were available from ALS patients, while 28 were from controls (176 samples in all) (Table S1A). Most of these patients presented with sporadic ALS disease (i.e., no known family history or pathogenic mutation) consistent with general estimates that ALS as a disease is largely sporadic (Chia et al., 2018).

The large size of this dataset enables *de novo* clustering algorithms to determine whether the ALS patient samples fall into distinct molecular subsets defined by specific gene signatures. We used a recent method for *de novo* clustering that was originally developed for single-cell sequencing datasets, but has been equally validated on bulk transcriptomes, single-cell RNA sequencing (RNA-seq) analysis and klustering evaluation (SAKE) (Ho et al., 2018). SAKE, developed by our group, is based on non-negative matrix factorization (NMF) and can robustly estimate both the number of clusters present in a given dataset and the confidence in assigning each sample to a cluster. SAKE returned optimal results for 3 clusters within the NYGC ALS Consortium Cohort (Figure S1), and a heatmap of the genes that define these clusters is given in Figure 1B. Based on the sets of gene markers returned by the SAKE algorithms, some of which are labeled on the left hand side of Figure 1B (see also Table S2A), we have chosen to re-label the 3 ALS subtypes as ALS-TE (elevated transposable element, or TE, expression, red

in Figure 1B), ALS-Ox (oxidative stress markers, blue), and ALS-Glia (elevated glial markers, gold).

The largest subset of patient samples in the NYGC survey (91/148) were marked by elevated levels of several genes previously associated with ALS. These included *NEFH*, *SOD1*, and *CDH13* (Chia et al., 2018), as well as a general expression signature consistent with oxidative and proteotoxic stress. These samples are identified as ALS-Ox (Figure 1B, blue). ALS-Ox patient samples showed elevated levels of genes in the oxidative phosphorylation pathway, in pathways associated with Parkinson's disease, as well as proteotoxic stress pathways as determined by Gene Set Enrichment Analysis (GSEA) (Figure 1C; Table S2B). Surprisingly, these samples also showed enrichment for genes previously noted to be elevated in both mutant SOD1 (Chiu et al., 2013) and TREM2-dependent (Keren-Shaul et al., 2017) models of disease associated microglia (Figure 1C), although microglial markers were not specifically enriched in these samples (Table S2B).

A second subset of patient samples, dubbed ALS-Glial (Figure 1B, gold), is defined by increased expression of genes that mark astrocytes (*CD44*, *GFAP*) and oligodendrocytes (*MOG*, *OLIG2*) suggesting that glial markers were a dominant signature in the transcriptomes of these samples. The ALS-Glial subset of patient samples also showed strong enrichment for transcriptional signatures previously identified as marking disease-associated microglia in either a mutant SOD1 (Chiu et al., 2013) or TREM2-dependent model (Keren-Shaul et al., 2017) (Figure 1C; Table S2C) and in levels of *TREM2* expression (Figure 1D). Finally, these patients also showed upregulation of innate immune pathways that are typically elevated in activated microglia, such as interferon and antigen processing pathways (Figure 1C).

Retrotransposon transcripts form a subset of the genes that define the remaining ALS samples, ALS-TE, named for transposable elements (TEs). These patient samples (Figure 1B, red) represented 20% (29/148) of the samples in the NYGC cohort. The ALS-TE samples showed an enrichment for transposon expression as the most significantly enriched pathway relative to controls using GSEA pathway analysis (Figure 1C; Table S2D), and this included TEs from the long interspersed element (LINE), short interspersed element (SINE), and long terminal repeat (LTR) classes. The samples in the ALS-TE subset also showed depletion for components of the spliceosome and for protein export pathways, two pathways previously linked to normal TDP-43 function (Cohen et al., 2011). Finally, violin plots of individual transposable elements, such as the LINE element L1PA6, show increased transposon levels for specific retrotransposons (Figure 1D). The fact that unbiased and unsupervised clustering identified several individual TEs as specifically defining this group provides a rigorous and quantitative method of determining the samples in which TE expression is well above the normal control levels and unique to this subset of ALS patients.

As stated above, multiple frontal and motor cortex tissues were sampled for a subset of ALS patients, allowing for an estimation of the level of concordance of ALS subtype between different tissues for the same patient. Of the 40 patients with both frontal and motor cortex samples sequenced, we note that only 7 were discordant between the two tissues (17.5%),
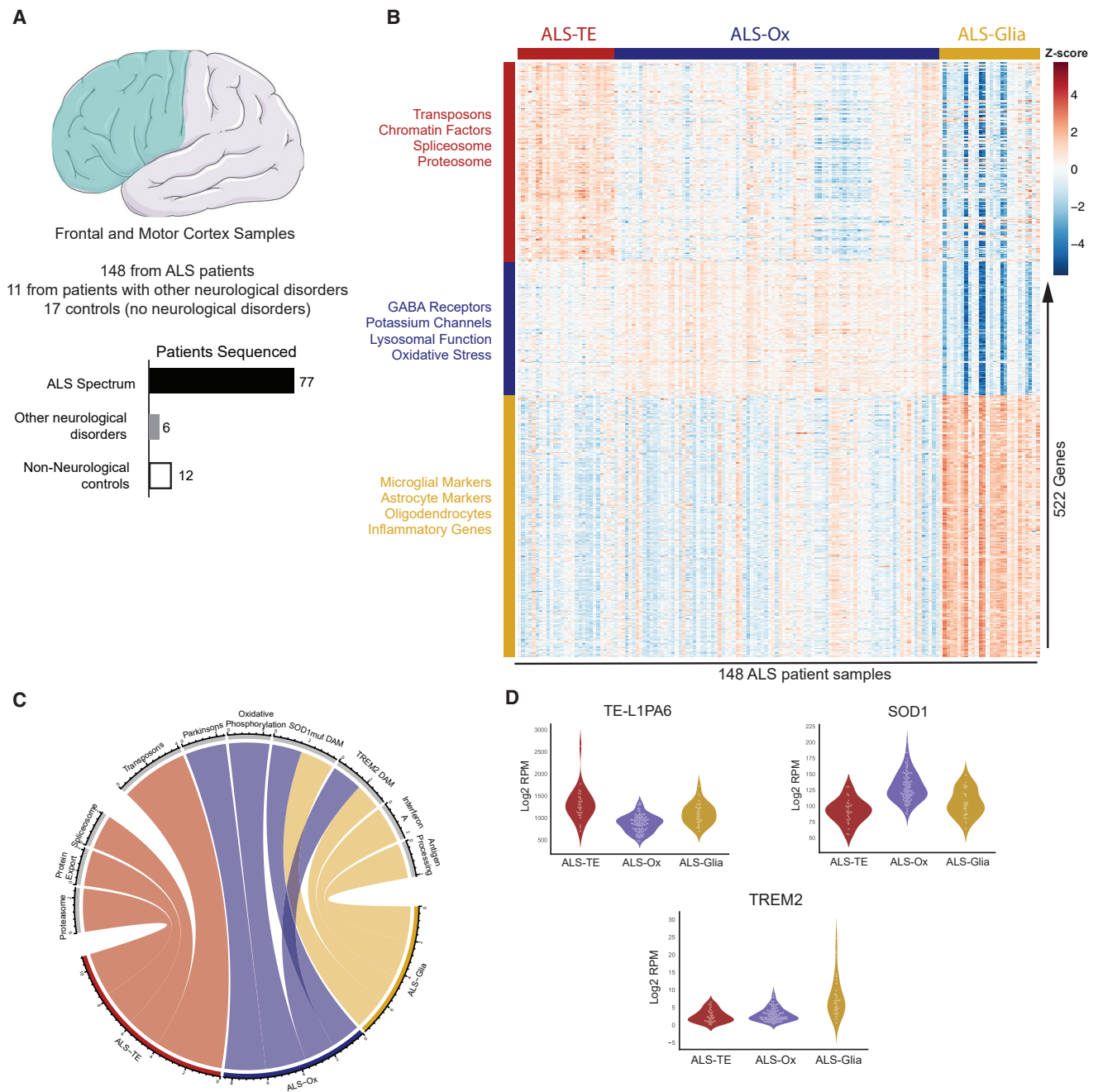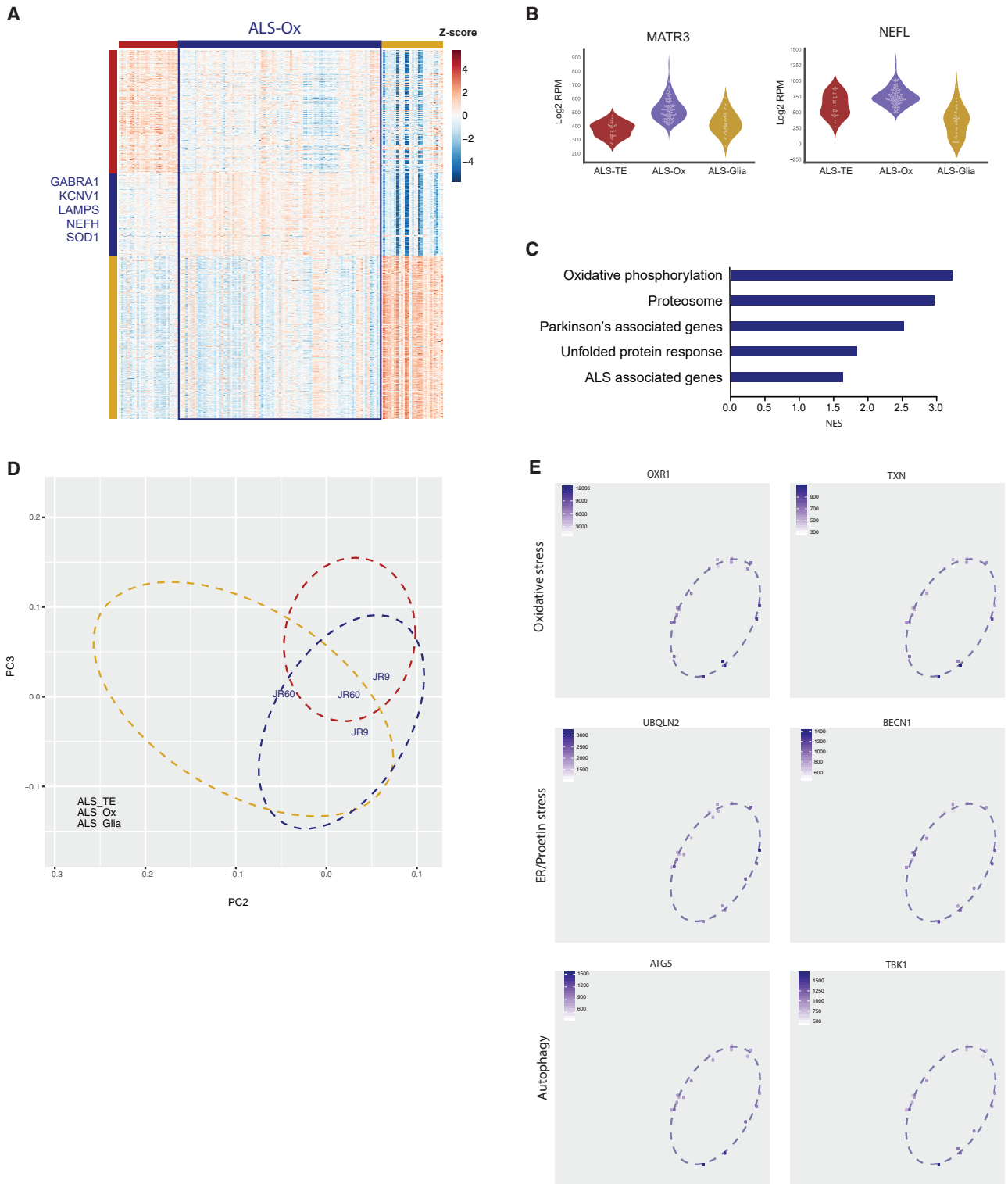
**Figure 1. ALS Patients in the NYGC ALS Consortium Cohort Show 3 Distinct Subtypes: Retrotransposon Reactivated (ALS-TE), Oxidative Stress Signatures (ALS-Ox), and Glial Dysfunction (ALS-Glial)**

(A) A schematic depiction of cortical regions from which post-mortem tissue samples were obtained: 148 from ALS patients, 11 from patients with other neurological disorders, and 17 from subjects. Multiple samples were obtained for most patients, such that these samples represent 77 ALS patients, 6 neurological control subjects, and 12 healthy controls.

(B) A heatmap of gene expression in ALS patient samples showing the results of NMF-based clustering, which returned 3 robust and distinct groups of patients. The retrotransposon reactivated set showed high TE expression as a dominant signature and formed its own subgroup (ALS-TE), representing 20% of all patient samples. Two additional groups were identified that showed signatures of oxidative stress (ALS-Ox, 61% of samples), and glial dysfunction (ALS-Glial, 19% of samples). Representative gene sets that describe markers from each subset are identified along the left of the heatmap.

(C) Gene Set Enrichment Analysis was performed for each group of patients (ALS-TE, ALS-Ox, and ALS-Glial) versus controls to identify group-specific aberrant pathways.

(D) Violin plots of example markers for each group show the LINE retrotransposon L1PA6 marking the ALS-TE group, *SOD1* marking the ALS-Ox group, and *TREM2* marking the ALS-Glial samples.

**Figure 2. ALS-Ox Patients in the NYGC ALS Consortium Show Markers of Oxidative Stress**

(A) The ALS-Ox group represents the middle 61% of patient samples in the ALS Consortium set (outlined in blue), with ALS-Ox marker genes listed along the left-hand side of the heatmap.

(B) Violin plots of example markers for this group show two known ALS-associated genes, *MATR3* and *NEFL*, specifically elevated in the ALS-Ox group (blue violins).

suggesting a high degree of overlap between these two tissues (Table S1B).

## The ALS-Ox Group Displays Evidence of Oxidative Stress

The importance of oxidative stress in neurodegenerative disease began with the initial identification of superoxide dismutase 1 (SOD1) as the first ALS-associated mutation (Rosen et al., 1993). Subsequent studies have identified a diverse array of functional pathways altered in SOD1 mutant mouse models, including SOD1 mediated autophagy (Rudnick et al., 2017), proteotoxic stress (Bruijn et al., 1998), and neuroinflammation (Chiu et al., 2013), reflecting both the pleiotropic roles played by SOD1 protein as well as different manifestations depending on the cellular context in which SOD1 is dysfunctional. Since that initial discovery, several genes with roles in oxidative stress, proteotoxic stress, and autophagy have been linked to ALS (Taylor et al., 2016). Consistent with the importance of these pathways, and their linked nature in generating neuronal stress, we found that 61% of the NYGC ALS patient samples displayed gene expression signatures consistent with a robust response to oxidative and proteotoxic stress, as described below.

Samples in the ALS-Ox group show elevated levels of several stress response genes, mutations in which have been linked to ALS, including SOD1 itself (Figures 1D and 2A), as well as the neurofilament protein NEFL and the nuclear matrix protein MATR3 (Figure 2B). More generally, the pathways that are elevated in the ALS-Ox samples relative to controls include genes involved in oxidative phosphorylation, proteasomal components, and genes involved in the unfolded protein response (Figure 2C). This subset of patients also displayed elevated expression of several genes that have been previously associated with multiple neurodegenerative diseases including both ALS and Parkinson's disease, such that these disease pathways are also significantly enriched in the ALS-Ox group relative to controls (Figure 2C; Table S2B).

To further validate these results, we obtained fresh frozen motor cortex tissue for an additional set of 13 ALS patients and 6 non-neurological controls, provided by a tissue bank at the University of California, San Diego (UCSD) (Table S1C). Two biological replicate transcriptomes were sequenced from each patient and control tissue sample to ensure the results would be robust to within-patient heterogeneity. We note that the RNA-seq libraries for the UCSD cohort were prepared at a different site, using a different library preparation protocol (please see Method Details), and yet these patient datasets displayed similar groupings to those from the NYGC cohort. We quantitatively assigned these samples to each of the ALS subtype groups, by combining the two groups (NYGC and UCSD) into a single principal components analysis (PCA) (Figure 2D). Patients were assigned to a subtype based on the distance to each cluster centroid on the PCA graph, with most samples falling within the 95% confidence interval ellipse defining the NYGC group (Figure S2A). Two of the 13 ALS patients from this UCSD cohort fell within the ALS-Ox group, with 2 samples plotted per patient to show concordance between biological replicates. Consistent with their classification, we noted that these patient samples also displayed similar expression patterns for all ALS-Ox marker genes, comparable to that seen in the NYGC samples (Figure S2B, gray bars). We next performed qPCR expression analysis on all UCSD patient samples for a selection of genes whose dysregulation characterized each of the three molecular subtypes (Figure S2C). While the number of patient samples here is small relative to the larger NYGC cohort, corresponding significant differences were validated for NEFL (p < 0.01) and ATG5 (p < 0.03) compared to controls. BECN1 showed milder differences that did not reach statistical significance compared to controls (p < 0.3) but shows a similar trend relative to other ALS cortex samples as that seen in the larger NYGC survey (Figure S2C).

The relative positions on the PCA plot in Figure 2D represent the expression of genes underlying the oxidative and proteotoxic stress responses in both patient cohorts. This is confirmed by the individual PCA plots of specific marker genes in Figure 2E, in which the color intensity of each dot indicates the expression level. For the oxidative stress pathway, expression is shown for oxidation resistance 1 (OXR1) and thioredoxin (TXN). For the proteotoxic stress response pathway, expression is shown for ubiquilin-2 (UBQLN2) and beclin-1 (BECN1). Finally, for markers of autophagy, expression is shown for autophagy gene 5 (ATG5) and TANK binding kinase (TBK1). We note that these 3 pathways are typically linked, such that both oxidative stress and proteotoxic stress have been noted to induce an autophagic response in neurodegenerative disease (Wong and Holzbaur, 2015), consistent with these pathways showing elevation in the same subset of samples. Thus, these results are consistent with ALS patient samples in the ALS-Ox group mounting a robust response to several neuronal stressors, which may be concurrent in the tissue.

## The Predominance of Glial Markers in the ALS-Glia Group

Several recent studies have noted the important role that glial cells play in neurodegenerative disease. Astrocytes, cells that support neuronal function by providing nutrients and removing waste, have previously been shown to secrete neurotoxic factors when expressing ALS-associated mutations (Di Giorgio et al., 2007; Nagai et al., 2007). Microglia, the innate immune cells of the CNS, have been shown to become activated in several neurodegenerative diseases (Deczkowska et al., 2018), setting off a neuroinflammatory cascade that eventually results in motor neuron cell death. In the ALS Consortium Cohort, 19% of the

---

(C) Top enriched pathways identified by GSEA in the ALS-Ox group samples compared to controls.

(D) Additional ALS patient samples from UCSD are assigned to the NYGC ALS subgroups by position on this PCA plot (UCSD samples marked by patient ID in blue).

(E) To confirm the link between ALS-Ox membership and stress pathways, individual PCA plots are shown where each sample is colored by the normalized expression of genes in the Oxidative Stress pathway (OXR1, TXN), ER-linked proteotoxic stress pathways (UBQLN2, BECN1), and downstream genes related to autophagy (ATG5, TBK1).
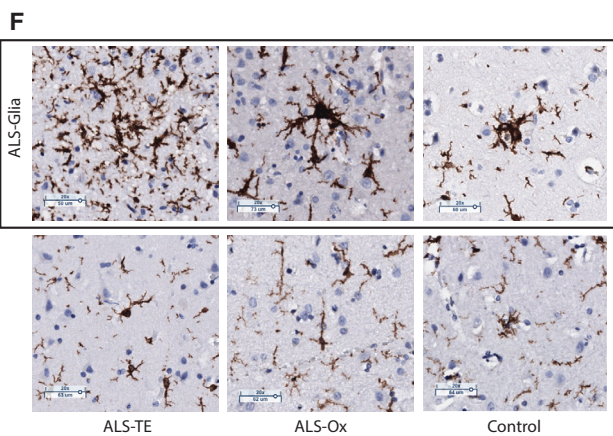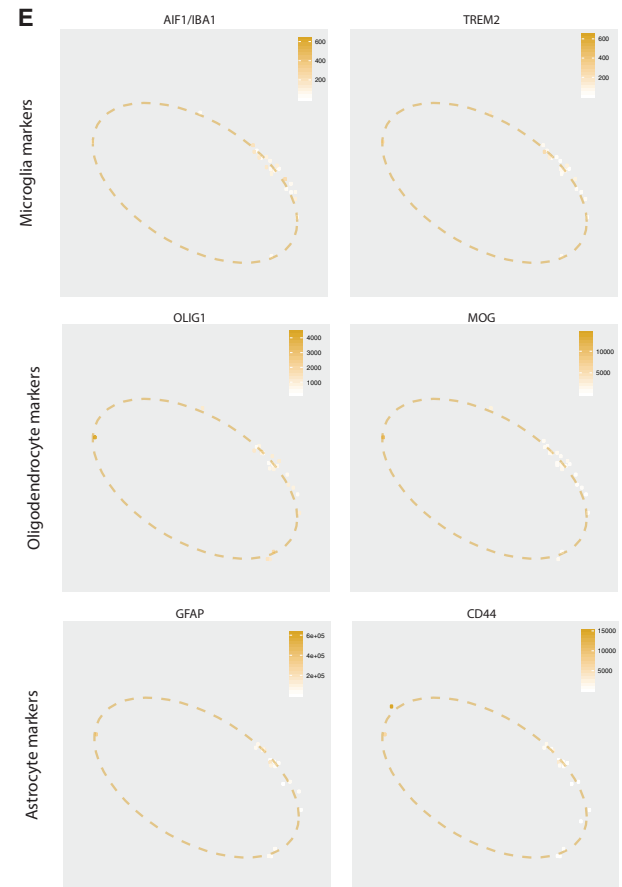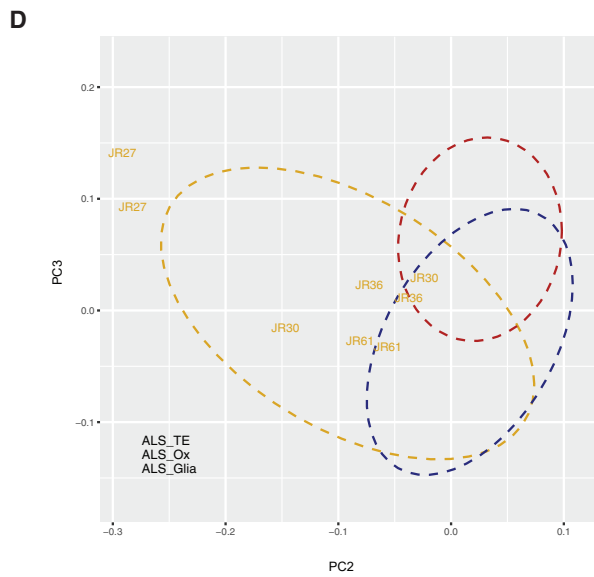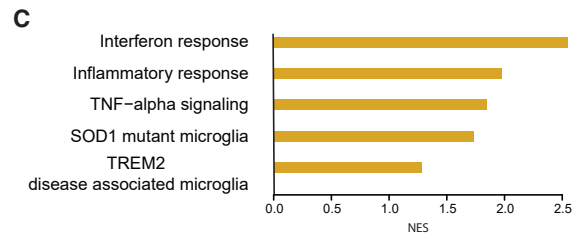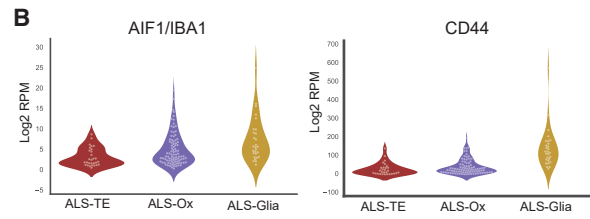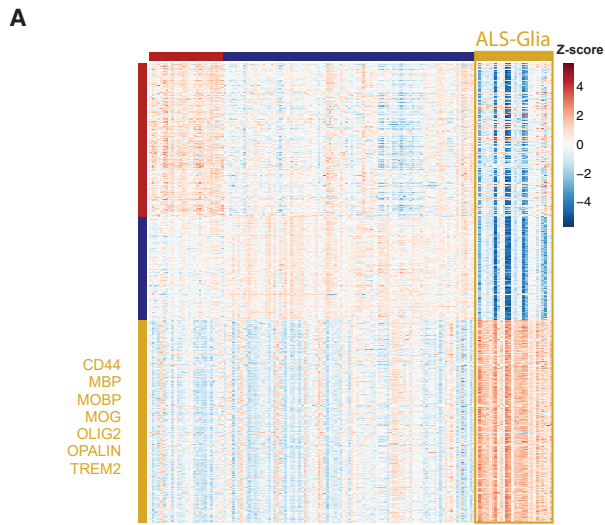
(legend on next page)

ALS patient samples showed extensive evidence of glial involvement, as discussed below.

Samples in the ALS-Glial subgroup show elevated expression of markers for all glial cell types, including astrocytes (*CD44*, *GFAP*), microglia (*IBA1/AIF1*, *TREM2*), and oligodendrocytes (*OLIG1*, *OLIG2*, *MOG*), as noted on the heatmap of Figure 3A. Violin plots are shown in Figure 3B for specific genes marking microglial (*IBA1*) and astrocyte (*CD44*) cell types in each of the ALS patient samples, with a particular enrichment in ALS-Glial subtype patients (gold). In particular, some of the microglial markers most prominently elevated in these samples represent genes known to be expressed in activated microglia and associated with neurodegenerative disease, such as *IBA1* (Figure 3B) and *TREM2* (Figure 1D). Consistent with this, pathway analysis of the genes that are upregulated in ALS-Glial samples relative to controls include mediators of the inflammatory response in general, the interferon response in particular, and other genes downstream of the tumor necrosis factor alpha (TNF-$\alpha$) signaling pathway (Figure 3C). Two recent reports have identified gene expression signatures of disease-associated microglia (DAM) in mouse models of neurodegeneration carrying mutations in either *SOD1* (Chiu et al., 2013) or *TREM2* (Keren-Shaul et al., 2017). These two DAM gene expression signatures were also strongly enriched in the ALS-Glia cortex samples relative to controls (Figure 3C; Table S2C).

One underlying mechanism that could explain profiles in the ALS-Glia samples is a relative loss of motor neurons, and/or increased presence of glial cells in these particular tissues. We used software designed to infer the relative composition of different cell types in bulk tissues, NeuroExpresso (Mancarci et al., 2017), to determine whether there was any evidence of differential cell type enrichment in the ALS-Glial samples. The NeuroExpresso results support this possibility, with estimates of cellular composition showing an increase in markers of activated microglia (p < 0.003) as well as oligodendrocytes (p < 0.002) and astrocytes (p < 0.02) (Figure S3A) relative to controls. These patients also showed the strongest signatures for relative loss of neuronal markers from both the pyramidal (p < 6.7e−7) and GABAergic (p < 3.1e−5) classes. We next obtained formalin-fixed paraffin-embedded (FFPE) slides matched to 35 of the decedents included in the NYGC cohort. These included 7 decedents from the ALS-Glia cluster, 9 decedents each from the ALS-TE and ALS-Ox clusters, and 6 controls. We performed immunohistochemistry for the microglial protein IBA1 (Figures 3F and S3B). While IBA1 marks all microglia (resting and activated),

we note that activated microglia are distinguished by more intense IBA1 staining and have enlarged cell bodies (diameters of $\sim$70 $\mu$m for activated versus $\sim$30 $\mu$m for resting, Figure S3B). Using these parameters, we quantified activated microglia in cortical layer sections. We find a spectrum in the rate of activated versus resting microglia in the cortex (Figure S3B) and found the highest rates of activated microglia in the cortex of decedent samples corresponding to the ALS-Glia cluster (Figure S3B). These moderate increases in activated microglia could be associated with neuronal cell loss in these tissues. To address this possibility, we measured cell density by counting all of the hematoxylin positive nuclei in each sample (Figure S3C). There were no significant differences between ALS subtypes and controls, and no decrease in cell number per mm$^2$ in the ALS-Glia patients, in particular. This suggests that the ALS-Glia transcriptome patterns are driven by activated transcription of glial cell associated pathways rather than relative loss of neuronal cells.

In the UCSD tissue validation cohort, 4/13 ALS patient samples clustered with the NYGC ALS-Glial group, without strong evidence of transposon de-silencing or oxidative stress markers (gold markers labeled by patient ID in the PCA plot of Figure 3D). These samples also showed elevated levels of TREM2 and IBA1 expression, at levels similar to that in the NYGC samples (Figure 3E). Moreover, markers of oligodendrocytes were also strongly enriched in this group, including OLIG1 and MOG. Finally, astrocyte markers were also present in these ALS-Glia samples, but were largely absent from samples in the ALS-TE and ALS-Ox groups, with GFAP and CD44 shown as representatives (Figure 3E). Validation by qPCR in the UCSD cohort, confirmed a significant enrichment for ALS-Glia markers in these patients relative to controls (Figure S2C): *TREM2* (p < 0.001), *MOBP* (p < 0.001), and *IBA1* (p < 0.01).

## The ALS-TE Group Displays Altered Expression of Genes in Multiple Pathways Linked to TDP-43

Sporadic ALS patients are known to show cytoplasmic accumulation and aggregation of TDP-43 protein (produced by the *TARDBP* gene) in the motor cortex and spinal cord, two tissues where motor neuron loss occurs (Arai et al., 2006; Neumann et al., 2006). Such TDP-43 pathology is thought to cause both a loss of the normal function of TDP-43 as well as aggregation associated detrimental impact. Thus, TDP-43 targets would be expected to be mis-regulated in ALS patients with TDP-43 pathology and an associated loss of TDP-43 nuclear function. Previous studies have linked TDP-43 to roles in splicing, mRNA

**Figure 3. ALS-Glia Patients in the NYGC ALS Consortium Cohort Show Evidence of Disease-Associated Microglia**
(A) The ALS-Glia group represents the 19% of patient samples on the right side of the ALS Consortium set (outlined in gold), with additional ALS-Glia marker genes listed along the left side of the heatmap.
(B) Violin plots of example markers for this group show two known glial marker genes, *AIF1/IBA1* and *CD44*, specifically elevated in the ALS-Glia group (gold violins).
(C) Top enriched pathways identified by GSEA in the ALS-Glia group samples compared to controls.
(D) Additional ALS patient samples from UCSD are assigned to the NYGC ALS subgroups by position on this PCA plot (UCSD samples marked by patient ID in gold).
(E) To confirm the link between ALS-Glia membership and markers of particular glial subsets, individual PCA plots are shown where each sample is colored by the normalized expression of genes that mark Microglia (*IBA1*, *TREM2*), Oligodendrocytes (*OLIG1*, *MOG*), and astrocytes (*GFAP*, *CD44*).
(F) To confirm the link between ALS-Glia membership and activated microglia, UCSD patient samples were stained with antibodies to IBA1. ALS-Glia samples show presence of enlarged, activated microglia not seen in controls or other ALS subgroups. Images are shown at 20 × magnification, with labeled scale bars indicating a size of 50–70$\mu$m.
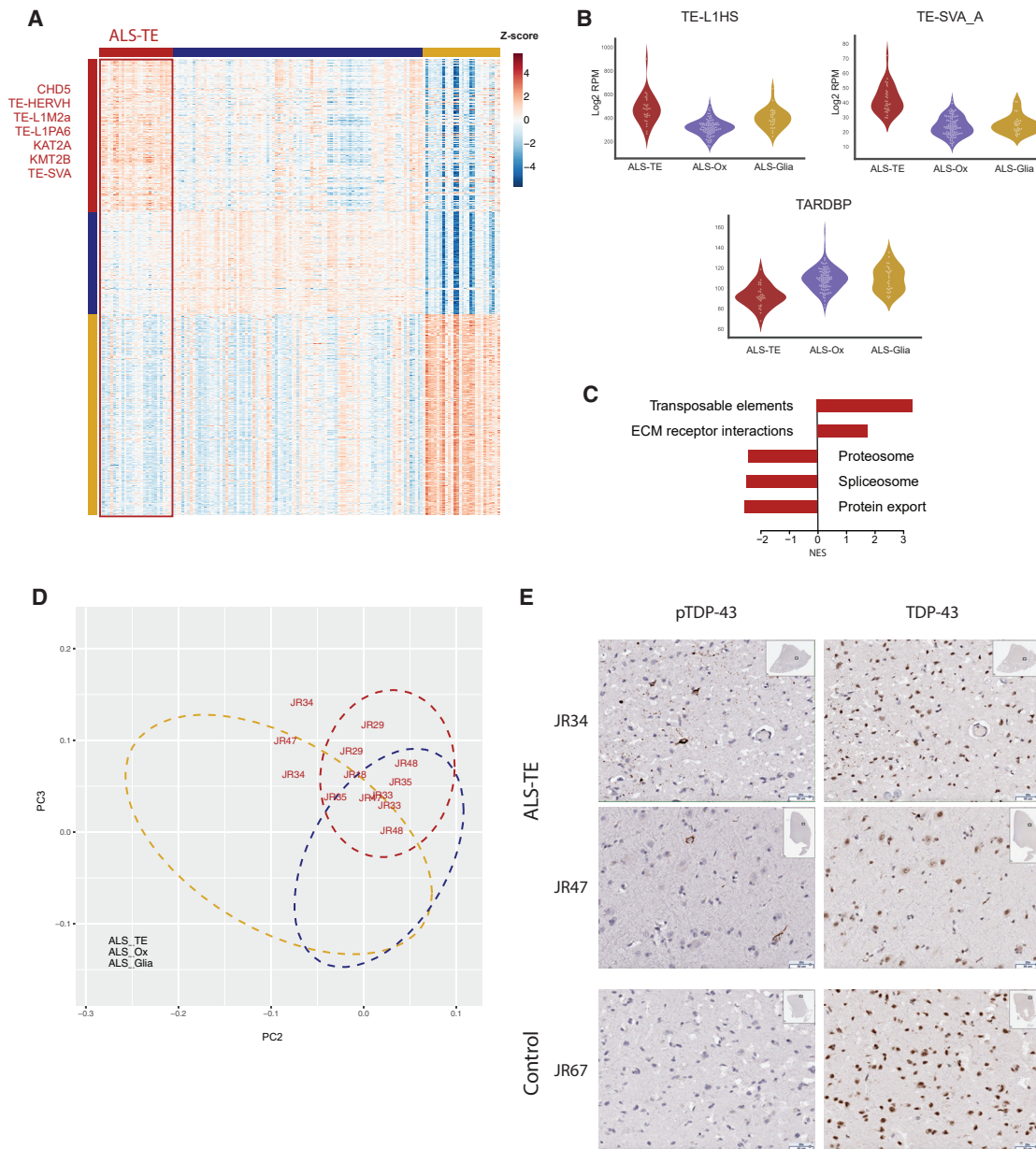
**Figure 4. ALS-TE Patients in the NYGC ALS Consortium Cohort Show Retrotransposon Re-activation**

(A) The ALS-TE group represents the left hand 20% of patients in the ALS Consortium sample set (outlined in Red), with ALS-TE markers listed along the left-hand panel.

(B) Violin plots of example markers for this group show two relatively young retrotransposons, L1HS and SVA_A, and TARDBP expression.

(C) Top enriched pathways identified by GSEA in the ALS-TE versus controls.

(D) Additional ALS samples from UCSD are assigned to the defined subgroups by position on this PCA plot (UCSD samples marked by patient ID in red).

(E) To confirm the link between ALS-TE membership and TDP-43 pathology, UCSD patient samples were stained with antibodies to the phosphorylated form of TDP-43 (pTDP-43, left) as well as with antibodies that recognize full-length TDP-43 protein (right). ALS-TE samples show evidence of pTDP-43 pathology not present in controls or other ALS subgroups. Images are shown at 20 × magnification, with labeled scale bars indicating a size of 50μm.

metabolism, and protein export (Cohen et al., 2011), as well as retrotransposon silencing (Krug et al., 2017; Li et al., 2012). Consistent with this last role, several retrotransposons from the endogenous retrovirus (ERV), LINE, and SINE-VNTR-*Alu* (SVA) class were selected as specifically marking the ALS-TE group

(Figure 4A). High levels of individual retrotransposons, including TEs from the LINE and SVA class (Figure 4B), characterize this group, while retrotransposons in general formed the top enriched pathway in GSEA (Figure 4C). Additional pathways depleted from the ALS-TE group are also consistent with

previously identified functional roles for TDP-43, including protein export, the spliceosome, and proteasomal components (Figure 4C). Finally, we note that *TARDBP* expression levels are the lowest in the ALS-TE group overall (Figure 4B), which could be mediated by TDP-43 auto-regulation (Ayala et al., 2011) or other mechanisms.

We again turned to samples from the UCSD patient cohort for follow up analysis and validation. Seven UCSD ALS decedents fell within the ALS-TE group, with patient ID marked in red on Figure 4D. Consistent with their PCA-based classification (Figure S2A), we noted that these UCSD patient samples also displayed elevated levels of TEs, comparable to that seen in the NYGC samples (Figure S2B). This was validated using qPCR for three specific retrotransposons that are known to be young and active in human genomes (Mills et al., 2007): the human specific LINE1 element (L1HS, p < 0.05) and two young SINE elements of the Alu class, AluYk12 (p < 0.001) and AluYa5 (p < 0.001) (Figure S2C).

To establish a more direct link between the ALS-TE group and TDP-43 pathology, we stained slides from both the NYGC and UCSD decedent samples described above with an antibody directed against phosphorylated TDP-43, which has been previously characterized to identify patients with TDP-43 pathology in ALS (Neumann et al., 2009). The ALS-TE samples with elevated TE expression levels and loss of known TDP-43 pathway genes are the most likely to show TDP-43 inclusion pathology in the frontal and motor cortex (p < 0.035), as shown for representative fields in Figure 4E and quantified in Figure S3D. Control samples did not show evidence of pTDP-43 inclusions in the tissues analyzed (Figure S3D), and non-ALS-TE samples did not exhibit transcriptional patterns consistent with established TDP-43 regulated pathways (Tables S2B and S2C). Together, these results establish a correlation between TDP-43 dysfunction and transcriptional re-activation of retrotransposon sequences in ALS patient tissues. The underlying basis for this will be explored next.

### TDP-43 Functions to Silence Retrotransposons

While pathways linked to oxidative stress and activated glia have previously been linked to ALS and other neurodegenerative diseases (Taylor et al., 2016), a potential role for retrotransposons in ALS is only recently emerging. As such, we used a genomics approach in neuronal-like cell lines to bolster the connection between retrotransposon expression and the function of the ALS associated protein TDP-43. TDP-43 is an RNA binding protein with two RNA recognition domains that recognize UGUGU repeat motifs present in thousands of cellular RNAs (Polymenidou et al., 2011; Tollervey et al., 2011). Previous TDP-43 studies using fly (Krug et al., 2017; Li et al., 2012) and human (Li et al., 2012) models to identify downstream targets of TDP-43 have suggested that retrotransposons may form a subset of TDP-43 targets, either directly or indirectly, but the extent and functional impact of this was not fully understood. To establish the subset of direct TDP-43 target genes and retrotransposons, we sequenced the RNAs bound to TDP-43 protein using an enhanced cross-linking and immunoprecipitation protocol (eCLIP-seq) (Van Nostrand et al., 2016) in human SH-SY5Y neuroblastoma cells. Peaks were called using a CLIP-seq analysis

tool designed for handling repetitive reads, CLAM (Zhang and Xing, 2017). Results from two biological replicates of TDP-43 eCLIP libraries were merged and normalized to a cross-linked, size-matched input control, resulting in 36,716 called peaks mapping to 5770 genes and 439 transposable elements (Figure 5; Tables S3A and S3B). Thirty-one percent of all peaks mapped to transposable elements generally (Figure 5B). However, 58% of the TE associated peaks mapped to the opposite strand, suggesting the TEs were providing regulatory sequence for the host gene, a phenomenon previously observed for other RNA binding proteins (Attig et al., 2018; Kelley et al., 2014; Zarnack et al., 2013). The remaining 42%, which were directly bound to TEs in the sense orientation, can be grouped into three major families, encompassing LINE elements (5% of all called peaks), SINE elements (4% of peaks), and LTR regions that include endogenous retroviruses (2%). Verification that these sequences were directly bound to TDP-43 protein was supported by the presence of TDP-43 binding motifs at called peaks in both the gene and retrotransposon annotated peaks (Figure 5C; Table S3C). Examples of TDP-43 CLIP reads over a known gene target (TARDBP/TDP-43) as well as retrotransposon targets from each major class (LINE, L1PA6; SINE, AluY; LTR, HERV3; SVA, SVA_D) can be seen in Figured 5A and 5D with a full list in Table S3B. The gene targets correlate well with previously identified TDP-43 targets from CLIP-seq and RIP-seq studies (Colombrita et al., 2012; Polymenidou et al., 2011; Tollervey et al., 2011; Van Nostrand et al., 2016; Xiao et al., 2011), which previously missed transposon targets due to masking of repetitive regions or discarding of ambiguously mapped reads. Together, these results demonstrate that TDP-43 normally binds broadly to both gene and TE-derived transcripts. This comprehensive list of direct targets in SH-SY5Y cells provides a platform to investigate the impact of *TARDBP* knock down on expression of target RNAs.

To determine the subset of CLIP targets that are regulated at the transcript abundance level, we knocked down the levels of TDP-43 protein using a short hairpin RNA strategy (TDP-43kd). The TDP-43kd construct strongly reduced *TARDBP* mRNA expression and TDP-43 protein level (Figure 5G). We sequenced the transcriptomes of both control and TDP-43kd libraries. All significantly altered retrotransposon transcripts were upregulated in TDP-43kd cells, indicating that TDP-43 normally contributes to the silencing of retrotransposon transcripts (Figure 5E; Table S3D). This is in contrast to altered gene transcripts, where a substantial fraction was either up- (1,165 genes, 34%) or downregulated (2,246 genes, 66%) in TDP-43kd cells (Figure 5F; Table S3D). Moreover, a majority of the directly bound gene targets were downregulated in the absence of TDP-43 (52%). This is consistent with known roles for TDP-43 in contributing to alternative splicing or translation control for a large fraction of gene targets that are not regulated at the mRNA abundance level (Polymenidou et al., 2011; Tollervey et al., 2011). However, the fact that nearly all expressed retrotransposon transcripts were upregulated in the absence of TDP-43 suggests that TDP-43 plays a silencing role for transposons. This is consistent with previous studies of the role of TDP-43 in the fly brain (Krug et al., 2017) and conclusively link directly bound TDP-43 transposon targets with the regulation of these targets at the expression level.
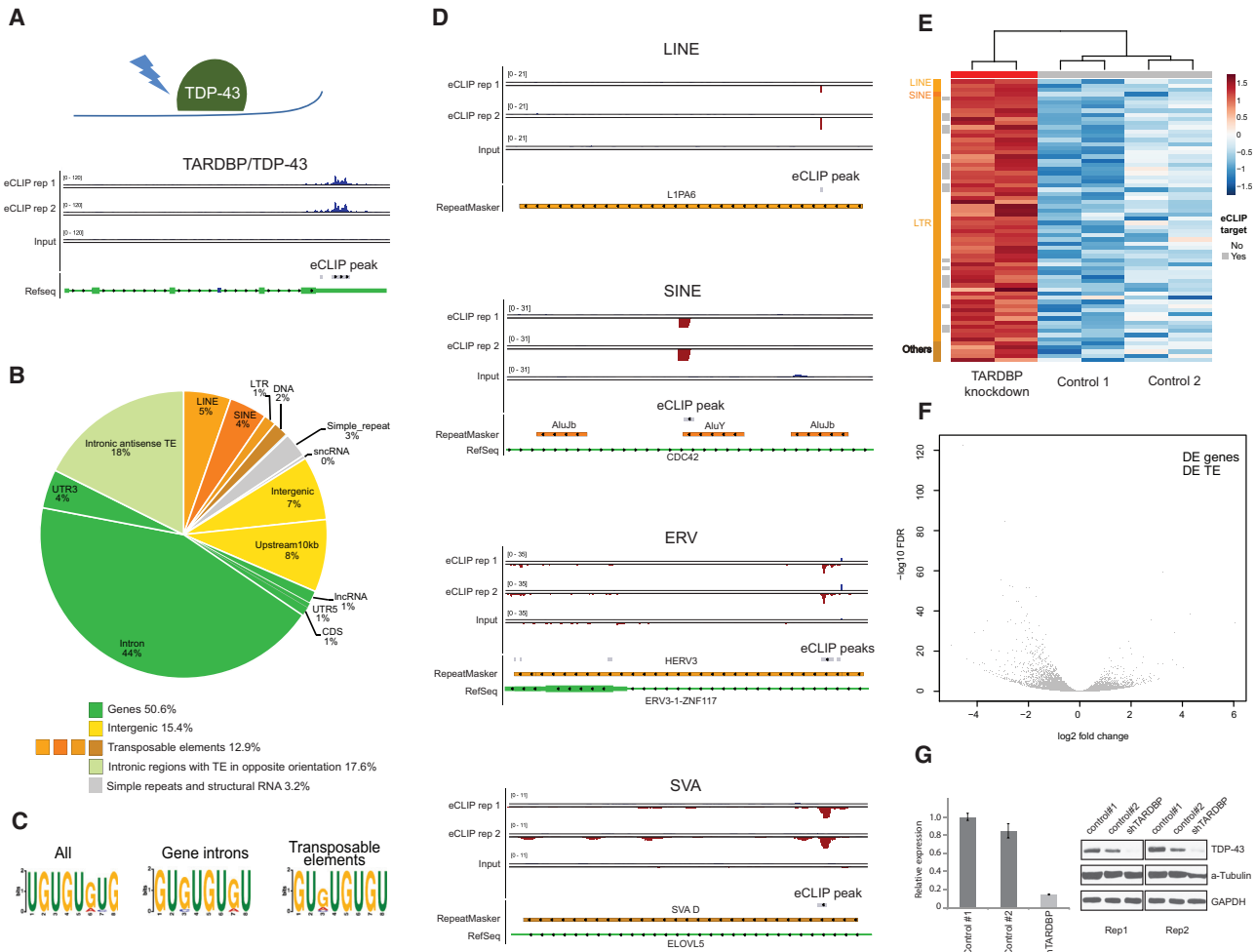
**Figure 5. TDP-43 Binds and Regulates Retrotransposon Transcripts**

(A) A schematic of the eCLIP-seq based identification (2 biological replicates) of directly bound RNA targets for TDP-43, with a representative pileup of reads where TDP-43 is known to target its own 3′ UTR.

(B) A pie chart of annotation categories for TDP-43 targets shows that retrotransposons are a small but important fraction of direct targets (12.9%). Gene intronic sequences represent the largest fraction of all peaks at 44%, while an additional 17.6% represent regions where TEs provide regulatory sequences to the host gene intron.

(C) The known TDP-43 binding motif (UGUGU) is present in both gene and retrotransposon targets.

(D) Representative browser pileup plots for one retrotransposon from each major repeat class: an L1PA6 LINE element, an AluY SINE element, a HERV3 LTR element, and an SVA_D from the human specific SVA class.

(E) Knock down of TDP-43 levels (2 biological replicates) results in the upregulation of retrotransposon transcripts, most of which also show evidence for direct binding in the eCLIP-seq experiments (gray bars).

(F) A volcano plot of fold change versus false discovery rate (FDR) q values shows that upregulated TEs (red dots) form a substantial fraction of all TDP-43 dependent genes, and all are upregulated.

(G) Knock down efficiency of short hairpin RNAs (shRNAs) targeting TARDBP was validated by western blot analysis and qPCR. For RNA levels, replicates were averaged and SD shown as error bars. For protein levels, western blots of the two biological replicates are shown separately.

## Clinical Parameter Association with ALS Subtype

We investigated whether clinical/phenotypic variables correlate with the three ALS molecular subtypes identified in this study. As shown in Figure S4, there was no statistically significant association between any of the groups and sex, C9ORF72 repeat expansion status, age at death, age at onset, or disease duration. Patients with limb onset were more frequently associated with the ALS-TE subtype than expected (Fisher's exact p <

0.03), however, there was no relative depletion of ALS-TE patient samples among those with bulbar onset symptoms. No significant differences in overall survival were found between the identified subtypes, as shown by Kaplan-Meier survival curves for each group (Figure S4G).

Genotyping was performed to assess the distribution of patients carrying germline repeat expansions in *C9orf72*. Hexanucleotide repeat expansions in the intron of *C9orf72* are the

most common pathogenic alterations associated with the development of ALS, both familial and sporadic (Majounie et al., 2012). Moreover, previous studies have noticed a slight trend toward higher TE expression in patients carrying *C9orf72* mutations (Pereira et al., 2018; Prudencio et al., 2017; Zhang et al., 2019). In order to determine if patients carrying a *C9orf72* repeat expansion are overrepresented in one or several of the identified ALS subtypes, genotyping of the hexanucleotide repeat size was performed on the NYGC cohort (Table S1B). *C9orf72* repeat expansion carriers were neither significantly over- nor underrepresented in the ALS-TE (p = 1) or ALS-Glia patients (p < 0.65). However, we observed a non-significant trend toward exclusion of *C9orf72* repeat expansion carriers among the ALS-Ox samples (p < 0.09). We note that none of the patient samples from the UCSD validation cohort carried repeat expansions in *C9orf72* (Table S1C), although each of the ALS subtypes was represented in this group (Figure S2B; Table S1C). Together, this suggests that *C9orf72* repeat expansions are not driving the transcriptional profiles of these subgroups but may be associated with patterns typically not seen in patients of the ALS-Ox group.

## DISCUSSION

*De novo* discovery of transcriptome subtypes from a large ALS cortex tissue sequencing study demonstrated that ALS cortex samples fall into three distinct clusters. The pathways that define two of these subtypes (oxidative and proteotoxic stress and neuroinflammation) have a well-established association with ALS disease, while a third subtype showed highly elevated retrotransposon expression. While genetic mutations are rare in ALS disease, pathogenic mutations can be grouped generally into those that largely mediate protein homeostasis, RNA metabolism, and neuroinflammation (Taylor et al., 2016), although many genes are linked to more than one of these processes. An example is *SOD1*, which has been ascribed roles in oxidative stress (Rosen et al., 1993), proteotoxic stress (Bruijn et al., 1998), and microglial activation (Chiu et al., 2013). While multiple cellular stressors might be contributing to the disease, we found that each decedent sample could be grouped into one of three molecular clusters, which largely reflected pathways already described for ALS, but which fell into surprisingly distinct groups. Specifically, 60% of the patient samples exhibited gene expression markers suggesting oxidative stress and proteotoxic stress were the main contributors to cellular dysfunction (the ALS-Ox group) but did not show elevated markers of glial cell types or inferred glial cell enrichment in the tissue downstream of these stress responses. Only 20% of the patients showed extensive glial involvement (the ALS-Glial group), with an unknown mechanism initiating the activation of glial cells for these patients in particular.

The role of transposable element expression is a relatively new topic in the study of neurodegeneration. Moreover, the connection between TDP-43 and transposable elements has only recently been explored. While previous studies have linked TDP-43 to transposon binding in human cells and to transposon regulation in animal models (Chang and Dubnau, 2019; Krug et al., 2017; Li et al., 2015; Liu et al., 2019), we show that TDP-

4-bound transposon transcripts are de-silenced in human cells and demonstrate that these same targets are de-silenced in a subset of human ALS patients. Additional mechanisms for transposon de-silencing in ALS patients with *C9orf72* repeat expansions have been suggested (Pereira et al., 2018; Prudencio et al., 2017; Zhang et al., 2019), although we note that *C9orf72* status was not strongly associated with the ALS-TE group in this study. Together, this suggests a model where retrotransposon silencing is a normal function of TDP-43 in somatic cells, and this role is disrupted in ALS patient tissues, potentially contributing to cellular toxicity. Previous studies have demonstrated that expression of envelope proteins from the endogenous retrovirus class of retrotransposons can be toxic to cells of the CNS (Antony et al., 2004; Kremer et al., 2013; Li et al., 2015). Additional studies have shown that expressed retrotransposon RNAs can also be toxic through aberrant recognition by innate immune components in Aicardi-Goutières syndrome (Crow and Manel, 2015). Other studies linked transposon activation in Alzheimer's disease with tau aggregation and suggest these alterations accompany neuroinflammation and genomic instability (Guo et al., 2018; Sun et al., 2018). DNA damage and structural variants induced by transposition itself are a formal possibility, given the elevated levels of the fully competent LINE-1Hs retrotransposon and evidence for active L1Hs transposition in human neurons (Evrony et al., 2016; Upton et al., 2015). However, most of the retrotransposons expressed in this study derived from fixed elements that have lost the capacity to transpose. These and other mechanisms for retrotransposon contributions to cellular toxicity remain to be explored in ALS patients but present a possible mechanism for cellular damage in the subset of patients with extensive TDP-43 pathology and retrotransposon re-activation.

Finally, the three subgroups defined in this study were identified from postmortem tissues of a largely sporadic set of patients with no known family history of the disease. Thus, ALS subtype appears to be largely independent from genotype. Transcriptional differences that separate these subtypes may reflect multiple contributing factors including causal mechanisms, disease progression stage, cellular composition of the tissues due to loss of particular cell types, and differences in environmental interactions that contribute to onset. The fact that transcriptional profiles of post mortem cortex tissues segregate into three distinct groupings may provide an entry point to investigate these underlying mechanisms. If peripheral tissues amenable to sampling or biopsy (e.g., peripheral blood monocytes, cerebrospinal fluid, or muscle) have distinct molecular signatures that correlate with patterns found in these CNS tissue subgroups, it could facilitate cohort selection for clinical trials of therapies targeting specific pathogenic mechanisms. Indeed, the existence of distinct clusters of sporadic ALS patients—whether based on intrinsic pathogenic differences or disease stage—might explain why many treatments identified in laboratory models targeting a specific pathway have failed to translate to successful clinical trials. Additional correlations may arise as the ALS Consortium genomic data cohort continues to grow. These additional studies will allow for an exploration of whether glial activation in the CNS correlates with general inflammation outside of CNS tissues.

OPEN
ACCESS
CellPress

# STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.09.066.

## CONSORTIA

The members of the NYGC ALS Consortium include Hemali Phatnani, Justin Kwan, Dhruv Sareen, James R. Broach, Zachary Simmons, Ximena Arcila-Londono, Edward B. Lee, Vivianna M. Van Deerlin, Neil A. Shneider, Ernest Fraenkel, Lyle W. Ostrow, Frank Baas, Noah Zaitlen, James D. Berry, Andrea Malaspina, Pietro Fratta, Gregory A. Cox, Leslie M. Thompson, Steve Finkbeiner, Efthimios Dardiotis, Timothy M. Miller, Siddharthan Chandran, Suvankar Pal, Eran Hornstein, Daniel J. MacGowan, Terry Heiman-Patterson, Molly G. Hammell, Nikolaos. A. Patsopoulos, Oleg Butovsky, Joshua Dubnau, Avindra Nath, Robert Bowser, Matt Harms, Eleonora Aronica, Mary Poss, Jennifer Phillips-Cremins, John Crary, Nazem Atassi, Dale J. Lange, Darius J. Adams, Leonidas Stefanis, Marc Gotkine, Robert Baloh, Suma Babu, Towfique Raj, Sabrina Paganoni, Ophir Shalem, Colin Smith, Bin Zhang, University of Maryland Brain and Tissue Bank, NIH NeuroBioBank, and Brent T. Harris. Additional information regarding the principal investigators of the NYGC ALS Consortium can be found in Document S2.

## REFERENCES

Antony, J.M., van Marle, G., Opii, W., Butterfield, D.A., Mallet, F., Yong, V.W., Wallace, J.L., Deacon, R.M., Warren, K., and Power, C. (2004). Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat. Neurosci. 7, 1088–1095.

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y., and Oda, T. (2006). TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochem. Biophys. Res. Commun. 351, 602–611.

Attig, J., Agostini, F., Gooding, C., Chakrabarti, A.M., Singh, A., Haberman, N., Zagalak, J.A., Emmett, W., Smith, C.W.J., Luscombe, N.M., and Ule, J. (2018). Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. Cell 174, 1067–1081.

Ayala, Y.M., De Conti, L., Avendaño-Vázquez, S.E., Dhir, A., Romano, M., D'Ambrogio, A., Tollervey, J., Ule, J., Baralle, M., Buratti, E., and Baralle, F.E. (2011). TDP-43 regulates its mRNA levels through a negative feedback loop. EMBO J. 30, 277–288.

Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., et al. (2017). QuPath: Open source software for digital pathology image analysis. Sci. Rep. 7, 16878.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11.

Bruijn, L.I., Houseweart, M.K., Kato, S., Anderson, K.L., Anderson, S.D., Ohama, E., Reaume, A.G., Scott, R.W., and Cleveland, D.W. (1998). Aggregation and motor neuron toxicity of an ALS-linked SOD1 mutant independent from wild-type SOD1. Science 281, 1851–1854.

Chang, Y.-H., and Dubnau, J. (2019). The Gypsy Endogenous Retrovirus Drives Non-Cell-Autonomous Propagation in a Drosophila TDP-43 Model of Neurodegeneration. Curr. Biol., S0960-9822(19)30951-0.

Chia, R., Chiò, A., and Traynor, B.J. (2018). Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. 17, 94–102.

Chiu, I.M., Morimoto, E.T.A., Goodarzi, H., Liao, J.T., O'Keeffe, S., Phatnani, H.P., Muratet, M., Carroll, M.C., Levy, S., Tavazoie, S., et al. (2013). A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model. Cell Rep. 4, 385–401.

Cohen, T.J., Lee, V.M.-Y., and Trojanowski, J.Q. (2011). TDP-43 functions and pathogenic mechanisms implicated in TDP-43 proteinopathies. Trends Mol. Med. 17, 659–667.

Colombrita, C., Onesto, E., Megiorni, F., Pizzuti, A., Baralle, F.E., Buratti, E., Silani, V., and Ratti, A. (2012). TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. J. Biol. Chem. 287, 15635–15647.

Crow, Y.J., and Manel, N. (2015). Aicardi-Goutières syndrome and the type I interferonopathies. Nat. Rev. Immunol. 15, 429–440.

Deczkowska, A., Keren-Shaul, H., Weiner, A., Colonna, M., Schwartz, M., and Amit, I. (2018). Disease-Associated Microglia: A Universal Immune Sensor of Neurodegeneration. Cell 173, 1073–1081.

Di Giorgio, F.P., Carrasco, M.A., Siao, M.C., Maniatis, T., and Eggan, K. (2007). Non-cell autonomous effect of glia on motor neurons in an embryonic stem cell-based ALS model. Nat. Neurosci. 10, 608–614.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Evrony, G.D., Lee, E., Park, P.J., and Walsh, C.A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. eLife 5, 56.

Guo, C., Jeong, H.-H., Hsieh, Y.-C., Klein, H.-U., Bennett, D.A., De Jager, P.L., Liu, Z., and Shulman, J.M. (2018). Tau Activates Transposable Elements in Alzheimer's Disease. Cell Rep. 23, 2874–2880.

Ho, Y.-J., Anaparthy, N., Molik, D., Mathew, G., Aicher, T., Patel, A., Hicks, J., and Hammell, M.G. (2018). Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. Genome Res. 28, 1353–1363.

Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics 31, 3593–3599.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kelley, D.R., Hendrickson, D.G., Tenen, D., and Rinn, J.L. (2014). Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. Genome Biol. 15, 537.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T.K., David, E., Baruch, K., Lara-Astaiso, D., Toth, B., et al. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. Cell 169, 1276–1290.

Kremer, D., Schichel, T., Förster, M., Tzekova, N., Bernard, C., van der Valk, P., van Horssen, J., Hartung, H.-P., Perron, H., and Küry, P. (2013). Human endogenous retrovirus type W envelope protein inhibits oligodendroglial precursor cell differentiation. Ann. Neurol. 74, 721–732.

Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.-W., Morrill, K., Prazak, L., Rozhkov, N., Theodorou, D., Hammell, M., and Dubnau, J. (2017).

Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLoS Genet. 13, e1006635.

Li, W., Jin, Y., Prazak, L., Hammell, M., and Dubnau, J. (2012). Transposable elements in TDP-43-mediated neurodegenerative disorders. PLoS ONE 7, e44099.

Li, W., Lee, M.H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., Campanac, E., Hoffman, D.A., von Geldern, G., Johnson, K., et al. (2015). Human endogenous retrovirus-K contributes to motor neuron disease. Sci. Transl. Med. 7, 307ra153.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 1, 417–425.

Liu, E.Y., Russ, J., Cali, C.P., Phan, J.M., Amlie-Wolf, A., and Lee, E.B. (2019). Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. Cell Rep. 27, 1409–1421.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

Macia, A., Widmann, T.J., Heras, S.R., Ayllon, V., Sanchez, L., Benkaddour-Boumzaouad, M., Muñoz-Lopez, M., Rubio, A., Amador-Cubero, S., Blanco-Jimenez, E., et al. (2017). Engineered LINE-1 retrotransposition in nondividing human neurons. Genome Res. 27, 335–348.

Majounie, E., Renton, A.E., Mok, K., Dopper, E.G.P., Waite, A., Rollinson, S., Chiò, A., Restagno, G., Nicolaou, N., Simon-Sanchez, J., et al.; Chromosome 9-ALS/FTD Consortium; French research network on FTLD/FTLD/ALS; ITALSGEN Consortium (2012). Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. Lancet Neurol. 11, 323–330.

Mancarci, B.O., Toker, L., Tripathy, S.J., Li, B., Rocco, B., Sibille, E., and Pavlidis, P. (2017). Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. eNeuro 4, ENEURO.0212-17.2017.

Mayer, J., Harz, C., Sanchez, L., Pereira, G.C., Maldener, E., Heras, S.R., Ostrow, L.W., Ravits, J., Batra, R., Meese, E., et al. (2018). Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins. Mol. Neurodegener. 13, 39.

Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. (2007). Which transposable elements are active in the human genome? Trends Genet. 23, 183–191.

Nagai, M., Re, D.B., Nagata, T., Chalazonitis, A., Jessell, T.M., Wichterle, H., and Przedborski, S. (2007). Astrocytes expressing ALS-linked mutated SOD1 release factors selectively toxic to motor neurons. Nat. Neurosci. 10, 615–622.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M., et al. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science 314, 130–133.

Neumann, M., Kwong, L.K., Lee, E.B., Kremmer, E., Flatley, A., Xu, Y., Forman, M.S., Troost, D., Kretzschmar, H.A., Trojanowski, J.Q., and Lee, V.M.-Y. (2009). Phosphorylation of S409/410 of TDP-43 is a consistent feature in all sporadic and familial forms of TDP-43 proteinopathies. Acta Neuropathol. 117, 137–149.

Nicolas, A., Kenna, K.P., Renton, A.E., Ticozzi, N., Faghri, F., Chia, R., Dominov, J.A., Kenna, B.J., Nalls, M.A., Keagle, P., et al.; ITALSGEN Consortium; Genomic Translation for ALS Care (GTAC) Consortium; ALS Sequencing Consortium; NYGC ALS Consortium; Answer ALS Foundation; Clinical Research in ALS and Related Disorders for Therapeutic Development (CReATe) Consortium; SLAGEN Consortium; French ALS Consortium; Project MinE ALS Sequencing Consortium (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron 97, 1268–1283.

Pereira, G.C., Sanchez, L., Schaughency, P.M., Rubio-Roldán, A., Choi, J.A., Planet, E., Batra, R., Turelli, P., Trono, D., Ostrow, L.W., et al. (2018). Properties of LINE-1 proteins and repeat element expression in the context of amyotrophic lateral sclerosis. Mob. DNA 9, 35.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.-C., Sun, E., Wancewicz, E., Mazur, C., et al. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat. Neurosci. 14, 459–468.

Prudencio, M., Gonzales, P.K., Cook, C.N., Gendron, T.F., Daughrity, L.M., Song, Y., Ebbert, M.T.W., van Blitterswijk, M., Zhang, Y.-J., Jansen-West, K., et al. (2017). Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. Hum. Mol. Genet. 26, 3421–3431.

Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., et al.; ITALSGEN Consortium (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72, 257–268.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. 29, 24–26.

Rosen, D.R., Siddique, T., Patterson, D., Figlewicz, D.A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J.P., Deng, H.X., et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362, 59–62.

Rudnick, N.D., Griffey, C.J., Guarnieri, P., Gerbino, V., Wang, X., Piersaint, J.A., Tapia, J.C., Rich, M.M., and Maniatis, T. (2017). Distinct roles for motor neuron autophagy early and late in the SOD1$^{G93A}$ mouse model of ALS. Proc. Natl. Acad. Sci. USA 114, E8294–E8303.

Saldi, T.K., Ash, P.E., Wilson, G., Gonzales, P., Garrido-Lecca, A., Roberts, C.M., Dostal, V., Gendron, T.F., Stein, L.D., Blumenthal, T., et al. (2014). TDP-1, the Caenorhabditis elegans ortholog of TDP-43, limits the accumulation of double-stranded RNA. EMBO J. 33, 2947–2966.

Spandidos, A., Wang, X., Wang, H., and Seed, B. (2010). PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. Nucleic Acids Res. 38, D792–D799.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550.

Sun, W., Samimi, H., Gamez, M., Zare, H., and Frost, B. (2018). Pathogenic tau-induced piRNA depletion promotes neuronal death through transposable element dysregulation in neurodegenerative tauopathies. Nat. Neurosci. 21, 1038–1048.

Taylor, J.P., Brown, R.H., Jr., and Cleveland, D.W. (2016). Decoding ALS: from genes to mechanism. Nature 539, 197–206.

Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A.L., Zupunski, V., et al. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat. Neurosci. 14, 452–458.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sánchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M., et al. (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell 161, 228–239.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods 13, 508–514.

van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., van der Spek, R.A.A., Võsa, U., de Jong, S., Robinson, M.R., et al.; PARALS Registry; SLALOM Group; SLAP Registry; FALS Sequencing Consortium; SLAGEN Consortium; NNIPPS Study Group (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat. Genet. 48, 1043–1048.

Wong, Y.C., and Holzbaur, E.L.F. (2015). Autophagosome dynamics in neurodegeneration at a glance. J. Cell Sci. 128, 1259–1267.

Xiao, S., Sanelli, T., Dib, S., Sheps, D., Findlater, J., Bilbao, J., Keith, J., Zinman, L., Rogaeva, E., and Robertson, J. (2011). RNA targets of TDP-43 identified by UV-CLIP are deregulated in ALS. Mol. Cell. Neurosci. 47, 167–180.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell 152, 453–466.

Zhang, Z., and Xing, Y. (2017). CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Res. 45, 9260–9271.

Zhang, Y.-J., Guo, L., Gonzales, P.K., Gendron, T.F., Wu, Y., Jansen-West, K., O'Raw, A.D., Pickles, S.R., Prudencio, M., Carlomagno, Y., et al. (2019). Heterochromatin anomalies and double-stranded RNA accumulation underlie C9orf72 poly(PR) toxicity. Science 363, eaav2606.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
| --- | --- | --- |
| Antibodies | | |
| Anti-GAPDH antibody | Cell Signaling Technology | Cat# 5174, RRID:AB_10622025 |
| Anti-IBA1 antibody | EnCor Biotechnology | Cat# RPCA-IBA1, RRID:AB_2722747 |
| Anti-phospho-TARDBP antibody (pS409/410) | Cosmo Bio Co., Ltd | Cat# TIP-PTD-M01, RRID:AB_1961900 |
| Anti-TARDBP antibody | Proteintech Group | Cat# 10782-2-AP, RRID:AB_615042 |
| Anti-alpha-tubulin antibody | CSHL monoclonal antibody collection | Clone DM1A, RRID:AB_5216186 |
| HRP-linked anti-mouse IgG | Cell Signaling Technology | Cat# 7076, RRID:AB_330924 |
| HRP-linked anti-rabbit IgG | Cell Signaling Technology | Cat# 7074, RRID:AB_2099233 |
| Biological Samples | | |
| Postmortem frontal and motor cortex samples from ALS patients and unaffected individuals | NYGC ALS Consortium | N/A |
| Postmortem motor cortex samples from ALS patients and unaffected individuals | UCSD | N/A |
| Critical Commercial Assays | | |
| AmplideX PCR/CE C9ORF72 Kit | Asuragen | Cat# 49581 |
| Bolt 4-12% Bis-Tris Plus Gel | ThermoFisher Scientific | Cat# NW04120BOX |
| PowerUp SYBR Green Master Mix | ThermoFisher Scientific | Cat# A25741 |
| SuperSignal West Pico PLUS Chemiluminescent Substrate | ThermoFisher Scientific | Cat# 34577 |
| Deposited Data | | |
| NYGC ALS Consortium RNA-seq datasets | GEO | GEO: GSE124439 |
| UCSD ALS RNA-seq datasets | GEO | GEO: GSE122650 |
| TARDBP knockdown in SH-SY5Y cell line | GEO | GEO: GSE122650 |
| TARDBP eCLIP in SH-SY5Y cell line | GEO | GEO: GSE122650 |
| Experimental Models: Cell Lines | | |
| SH-SY5Y cell line | ATCC | Cat# CRL-2266, RRID:CVCL_0019 |
| HEK293FT cell line | ThermoFisher Scientific | Cat# R70007, RRID:CVCL_6911 |
| Oligonucleotides | | |
| See Table S4 | | N/A |
| Recombinant DNA | | |
| MISSION® pLKO.1-puro Human TDP-43/TARDBP shRNA plasmid | Genetic Perturbation Platform | ID# TRCN0000016038 |
| MISSION® pLKO.1-puro Luciferase shRNA Control Plasmid | Sigma Aldrich | Cat# SHC007 |
| MISSION® pLKO.1-puro Non-Target shRNA Control Plasmid DNA | Sigma Aldrich | Cat# SHC016 |
| psPAX2 | Gift from Didier Trono | RRID:Addgene_12260 |
| pMD2.G | Gift from Didier Trono | RRID:Addgene_12259 |
| Software and Algorithms | | |
| CLAM v1.1.3 | Zhang and Xing, 2017 | https://github.com/Xinglab/CLAM |
| DESeq2 v1.22.2 | Love et al., 2014 | RRID:SCR_015687 |
| GSEA v3.0 | Subramanian et al., 2005 | RRID:SCR_003199 |
| QuPath | Bankhead et al., 2017 | https://qupath.github.io/ |
| SAKE v1.0 | Ho et al., 2018 | https://github.com/naikai/sake |
| STAR v2.5.2b | Dobin et al., 2013 | RRID:SCR_015899 |
| TEtranscripts (TEToolkit) v2.0.3 | Jin et al., 2015 | https://github.com/mhammell-laboratory/tetoolkit |

## LEAD CONTACT AND MATERIALS AVAILABILITY

Immediate access to new and ongoing data generated by the NYGC ALS Consortium and obtained from samples collected through the Target ALS post-mortem core can be requested at ALSData@nygenome.org. All RNA-seq data in the NYGC ALS Consortium are made immediately available to all members of the Consortium and with other Consortia with whom we have a reciprocal sharing arrangement. Further information and requests for resources and reagents generated outside of the NYGC ALS Consortium should be directed to Molly Gale Hammell (mhammell@cshl.edu).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### ALS Post-mortem Samples

The NYGC ALS Consortium samples in this study, the majority provided by the Target ALS Post-Mortem Tissue Core, were acquired through various IRB protocols from member sites and transferred to NYGC in accordance with all applicable foreign, domestic, federal, state, and local laws and regulations for processing, sequencing, and analyses. All available de-identified clinical and pathological records were collected and used together with C9orf72 genotypes to summarize patient demographics and disease features (see Tables S1A and S1B).

UCSD samples were obtained from patients who had been followed during the clinical course of their illness and met El Escorial criteria for definite ALS. These individuals had bulbar or arm onset of disease and caudally progressing disease. Control nervous systems were from patients from the hospital's critical care unit when life support was withdrawn. IRB approval was obtained from the University of California, San Diego Human Research Protection Program. All available de-identified patient data is available in Table S1C.

### Cell Line

SH-SY5Y cells (CRL-2266, ATCC, Manassas, VA, USA) were grown in DMEM/F-12 Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (11320033, ThermoFisher Scientific, Waltham, MA, USA) supplemented with 10% FBS and 1% penicillin–streptomycin, and cultured at 37°C with 5% $CO_2$.

HEK293FT cells (R70007, ThermoFisher Scientific, Waltham, MA, USA) were grown in DMEM medium containing 10% FBS, and cultured at 37°C with 5% $CO_2$.

## METHOD DETAILS

### Generation of RNA-seq Libraries from NYGC ALS Samples

RNA was extracted from flash-frozen patient samples homogenized in Trizol (15596026, ThermoFisher Scientific, Waltham, MA, USA) -Chloroform and purified using the QIAGEN RNeasy Mini kit (74104, QIAGEN, Germantown, MD, USA). RNA was assessed using the Bioanalyzer (G2939BA, Agilent, Santa Clara, CA, USA). RNA-seq libraries were prepared from 500 ng of total RNA using the KAPA Stranded RNA-seq kit with RiboErase (07962304001, Kapa Biosystems, Wilmington, MA, USA) for rRNA depletion and Illumina compatible indexes (NEXTflex RNA-seq Barcodes, NOVA-512915, PerkinElmer, Waltham, MA, USA). Pooled libraries (average insert size: 375bp) were sequenced on an Illumina HiSeq 2500 using a paired end 125 nucleotide setting, to yield ~40-50 million reads per library.

### Generation of RNA-seq Libraries from UCSD ALS Samples

RNA was extracted from each of the flash-frozen patient samples using the Ambion PureLink RNA Mini kit (12183020, ThermoFisher Scientific, Waltham, MA, USA). RNA was assessed using the Bioanalyzer to ensure RNA Integrity (RIN) values of > = 5.5. RNA-seq libraries were prepared from 500 ng of total RNA using the Illumina TruSeq Stranded Total RNA kit (20020596, Illumina, San Diego, CA, USA). Samples were barcoded to multiplex 8 samples per batch, randomly mixing ALS and control samples in each library preparation and sequencing batch. The libraries were sequenced on an Illumina NextSeq using a single end 76 nucleotide setting, and pooled to ensure ~40 million reads per library.

### shRNA Knockdown and Assessment of Efficiency

The MISSION pLKO.1-puro human TDP-43 (TRCN0000016038, Genetic Perturbation Platform) and control shRNAs, SHC007 luciferase shRNA and SHC016 Non-target shRNA control (Sigma Aldrich, St. Louis, MO, USA), were used to produce lentivirus. The pLKO.1 plasmid DNA, together with psPAX2 packaging and pMD2.G envelope plasmid DNA (Didier Trono lab, Addgene plasmid # 12260 and 12259 respectively) were combined at a ratio of 4:3:1, respectively and transfected with PEI reagent (23966-1, Polysciences, Warrington, PA, USA) into HEK293FT cells. Virus containing medium was collected 48 and 72 hours post-transfection and stored at −80°C.

One and a half to two million SH-SY5Y cells per well of a 6-well plate were "spinfected" with 3 mL of virus containing media and 8µg/mL polybrene for 1 hour at 800 g. Cells stably expressing shRNAs were selected with 2 µg/mL of puromycin 48 hours post-infection for 3 days. Uninfected cells at similar density were used as a control for puromycin selection. Cells were harvested 5 days after infection for downstream analyses.

TDP-43 protein level was assessed by western blot analysis on two biological replicates of SH-SY5Y cells treated with the TDP-43kd construct or control shRNAs (SHC007 luciferase shRNA or SHC016 Non-target shRNA control) as described above. In brief, 15 µg of protein was electrophoresed on a Bolt 4%–12% Bis-Tris Plus Gel (NW04120BOX, ThermoFisher Scientific, Waltham, MA, USA), and transferred to nitrocellulose membrane. Primary antibodies specific for TDP-43 (10782-2-AP, Proteintech, Rosemont, IL, USA, 1:3,000), alpha-tubulin (Clone DM1A, provided by the CSHL monoclonal antibody collection, Cold Spring Harbor, NY, USA, 1:3000) and GAPDH (5174, Cell Signaling Technology, Danvers, MA, USA, 1:1000) were used. Protein detection was performed using HRP-linked secondary antibodies (7074 and 7076, Cell Signaling Technology, Danvers, MA, USA, 1:10000), and Super Signal West Pico PLUS Chemiluminescent Substrate (34577, ThermoFisher Scientific, Waltham, MA, USA).

Knockdown efficiency of *TARDBP* transcript was assessed by quantitative PCR, performed in triplicate on two biological replicates of SH-SY5Y cells treated with the TDP-43kd construct or control shRNAs (SHC007 luciferase shRNA or SHC016 Non-target shRNA control) as described above. Total RNA was extracted using Trizol according to the manufacturer's instructions from the harvested cells. qPCR reactions were made with the PowerUp SYBR Green Master Mix according to the manufacturers' instructions (A25741, ThermoFisher Scientific, Waltham, MA, USA), and run on a QuantStudio 6 Flex instrument (ThermoFisher Scientific, Waltham, MA, USA). Primers for TARDBP and GAPDH were obtained from PrimerBase (Spandidos et al., 2010). Please see Table S4 for more details on the primers used.

### RNA-seq in SH-SY5Y Cells

Total RNA was extracted using Trizol according to the manufacturer's instructions from two biological replicates of SH-SY5Y cells treated with the TDP-43kd construct or control shRNAs (SHC007 luciferase shRNA or SHC016 Non-target shRNA control) as described above. One microgram of total RNA was subjected to rRNA removal using the Ribo-Zero Gold rRNA Removal kit (MRZG126, Illumina, San Diego, CA, USA). Strand specific RNA-seq libraries were constructed using NEBNext Ultra II Directional Library Prep Kit (E7760S, New England Biosciences, Ipswich, MA, USA). The libraries were sequenced on an Illumina NextSeq using a single end 76 nucleotide setting, and pooled to ensure ∼40 million reads per library.

### TARDBP/TDP-43 eCLIP in SH-SY5Y Cells

Twenty million SH-SY5Y cells were UV-crosslinked (254 nm, 400 mJ/cm$^2$) in PBS and cell pellets were stored at −80°C. Each sample was prepared as an independent biological replicate and included non-crosslinked cells as a control. Cell pellets were lysed in 1 mL lysis buffer, followed by RNase I digestion and immunoprecipitation with 10 µg of anti-TDP-43 antibody (10782-2-AP, Proteintech, Rosemont, IL, USA).

eCLIP libraries were prepared as previously described (Van Nostrand et al., 2016). In brief, a barcoded adaptor is ligated to the 3′ end, and the precipitated RNA-protein complex is separated on nitrocellulose. The complex is digested with proteinase K and the RNA is precipitated and reverse-transcribed using SuperScript IV first strand synthesis system (18091050, ThermoFisher Scientific, Waltham, MA, USA). A second adaptor is ligated on the 5′ end, and the libraries were sequenced on an Illumina NextSeq500 using a single end 76 nucleotide setting.

### Histology and Immunohistochemistry

Frozen motor cortex tissues were fixed in 10% neutral buffered formalin and 5 µm paraffin sections were cut for immunohistochemistry (IHC) staining. IHC staining was performed on Ventana Discovery Ultra platform with OmniMap HRP and ChromoMap DAB detection system according to manufacturer's protocols (Roche, Indianapolis, IN, USA), using primary antibodies specific for TDP-43 (10782-2-AP, Proteintech, Rosemont, IL, USA, 1:10,000), phosphoTDP-43 (pS409/410) (CAC-TIP-PTD-M01, Cosmo Bio USA, Co., Carlsbad, CA, 1:200) and IBA1 (RPCA-IBA1, EnCor Biotechnology Inc., Gainesville, FL, 1:2,000). Slides were counterstained with Hematoxylin, and then scanned by Leica Aperio ScanScope system.

### Genotyping of C9orf72 Repeat Expansion

NYGC patient samples were genotyped for *C9orf72* repeat expansions using a combination of cross-repeat fragment size analysis, standard repeat-primed PCR, and the Asuragen AmplideX PCR/CE C9ORF72 Kit (49581, Asuragen, Austin, TX, USA).

UCSD patient samples were genotyped for *C9orf72* repeat expansions using standard repeat-primed PCR, as previously described (Renton et al., 2011). Briefly, primers outside the *C9orf72* intronic repeats were amplified in a nested PCR strategy, and the resulting traces were analyzed on a Bioanalyzer to determine repeat length.

### qPCR Analysis of UCSD Patient Samples

Quantitative PCR of UCSD samples was carried out in triplicate on two biological replicates each, using PowerUp SYBR Green Master Mix according to the manufacturers' instructions (A25741, ThermoFisher Scientific, Waltham, MA, USA), and run on a QuantStudio 6 Flex instrument (ThermoFisher Scientific, Waltham, MA, USA). Primers for *AIF1/IBA1*, *ATG5*, *BECN1*, *GAPDH*,

*MOBP*, *NEFL* and *TREM2* were obtained from PrimerBank (Spandidos et al., 2010). Primers for RPS24 were obtained from Sigma Aldrich (H_RPS24_1, Sigma Aldrich, St. Louis, MO, USA). Previously published primers for L1HS 5′ UTR (Macia et al., 2017), AluYa5 and AluYk12 (Prudencio et al., 2017) were used for quantification of these transposable elements.

Please refer to Table S4 for more details on the primers used.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Analysis of RNA-seq Libraries
Reads from samples with RIN > = 5.5 were aligned to the hg19 human genome using STAR v2.5.2b (Dobin et al., 2013), allowing for a 4% mismatch rate and up to 100 alignments per read to ensure capture of young transposon sequences. Abundance of gene and transposon sequences was calculated with TEtranscripts v2.0.3 (Jin et al., 2015). For differential expression analysis, we employed DESeq2 (Love et al., 2014), using the DESeq normalization strategy and negative binomial modeling. B-H corrected FDR P value threshold of $p < 0.05$ was used to determine significance. For heatmap visualization, the reads were normalized using a variance stabilizing transformation in DESeq2.

### Transcriptome *De Novo* Cluster Identification
The number of clusters in the ALS datasets was determined using the SAKE software suite (Ho et al., 2018). Briefly, a variance stabilizing transformation was performed on the raw counts data using DESeq2. Gender associated genes were then removed from this list before rank ordering by median absolute deviation using the SAKE software suite, selecting the top 5000 genes with the highest median absolute deviation (MAD). SAKE was implemented using 200 iterations, the "nsNMF" algorithm for Non-negative Matrix Factorization, and a "k" setting of 3 clusters for the cortex samples, determined by the k setting with the highest cophenetic correlation coefficient. Wilcoxon Mann Whitney U-tests were used to determine differences between ALS Subgroups and correlations with gene expression or cell type composition.

### Classification of UCSD Samples using NYGC-identified Clusters
Variance stabilizing transformation was performed on the raw counts of NYGC and UCSD data using DESeq2. Principal component analysis was then performed to identify principal components that delineate the identified ALS subtypes, and a 95% confidence ellipse was calculated from the NYGC samples for each subtype. The position of the UCSD patient samples along the principal components of interest was determined as the midpoint of the two replicates, and a distance to the center of the 95% confidence ellipse for each subtype was calculated. The UCSD samples were classified based on the nearest ellipse center of the subtype.

### Gene Set Enrichment Analysis
Pre-generated gene sets used for GSEA (Subramanian et al., 2005) were obtained from the MSigDB Hallmark collection (Liberzon et al., 2015), and KEGG (Kanehisa and Goto, 2000). Custom gene sets were generated from the SOD1 (Chiu et al., 2013) and TREM2 (Keren-Shaul et al., 2017) models of disease-associated microglia, as well as TARDBP targets from this study. Custom transposable element sets, including retrotransposons considered active in the human genome (Mills et al., 2007) were generated from Repbase (Bao et al., 2015).

### Analysis of eCLIP Libraries
Reads were trimmed to remove adapters, and aligned to the hg19 human genome using STAR v2.5.2b (Dobin et al., 2013), allowing for a 4% mismatch rate and up to 100 alignments per read to ensure capture of young transposon sequences. Weighting of multi-mapper alignments (expectation-maximization modeling) and identification of regions of enrichment relative to input (using negative binomial modeling) were performed with CLAM v1.1.3 (Zhang and Xing, 2017). Detected sites with B-H corrected FDR P value of $p < 0.05$ were considered to be significantly enriched. Integrated genome viewer (Robinson et al., 2011) was used for visualization of enriched regions and read depth.

### Analysis of IHC Tissue Slides
Images were exported from the Leica Aperio ScanScope and analyzed using the QuPath image analysis software v 0.2.0-m2 (Bankhead et al., 2017). Upper cortical regions within 1.5mm of the outer cortical surface of the tissue that were of good quality (minimal damage or non-specific staining) were chosen as regions of interest. Cell nuclei were detected in the regions of interest using Hematoxylin optical density (minimum nuclear area of $15\mu m^2$, but otherwise default settings). Cells positive for phosphorylated TDP-43 accumulation were identified as having mean nuclear DAB optical density above 0.75. Activated microglia were designated as cells with nuclear area of at least $70\mu m^2$, and a mean nuclear DAB optical density above 0.7. Proportions of positive cells (pTDP-43 or activated microglia) were calculated. Cell density in regions of interest was determined by dividing the number of detected nuclei by the area of the regions of interest. Wilcoxon Mann Whitney U-tests were used for comparison of pTDP-43/IBA1 positive cell proportions and cell density between ALS Subgroups and control samples.

### qPCR Data Analysis

Quantitative PCR results of UCSD patient samples were analyzed using the ΔΔCt method, normalized to *GAPDH* and *RPS24*. Knock-down efficiency of TDP-43kd construct was determined using the ΔΔCt method, normalized to *GAPDH*. *TARDBP* transcript levels are expressed relative to the SHC007 transfected samples. A Student's t test was used to determine statistical significance.

### Analysis of Clinical Parameters

Fisher's Exact test was used to determine enrichments of patient gender and *C9orf72* repeat expansion in each ALS subtype. Differences in age of death (compared to control individuals), age of onset and disease duration (compared between subtypes) was analyzed using Wilcoxon Mann Whitney U-tests. Survival analysis was performed using Kaplan-Meier log rank test.

### DATA AND CODE AVAILABILITY

The accession number for the CSHL Motor Cortex RNA-seq datasets reported in this paper is GEO: GSE122649. The accession number for the CLIP-seq and RNA-seq datasets from SH-SY5Y cells reported in this paper is GEO: GSE122650. The accession number for the RNA-seq datasets generated by the NYGC ALS Consortium reported in this paper is GEO: GSE124439.