HHS Public Access

Author manuscript

*Sci Transl Med*. Author manuscript; available in PMC 2024 August 15.

Published in final edited form as:

*Sci Transl Med*. 2024 February 14; 16(734): eadg7162. doi:10.1126/scitranslmed.adg7162.

# Mis-spliced transcripts generate de novo proteins in TDP-43–related ALS/FTD

*A full list of authors and affiliations appears at the end of the article.*

## Abstract

Functional loss of TDP-43, an RNA binding protein genetically and pathologically linked to amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), leads to the inclusion of cryptic exons in hundreds of transcripts during disease. Cryptic exons can promote the degradation of affected transcripts, deleteriously altering cellular function through loss-of-function mechanisms. Here, we show that mRNA transcripts harboring cryptic exons generated de novo proteins in TDP-43–depleted human iPSC–derived neurons in vitro, and de novo peptides were found in cerebrospinal fluid (CSF) samples from patients with ALS or FTD. Using coordinated transcriptomic and proteomic studies of TDP-43–depleted human iPSC–derived neurons, we identified 65 peptides that mapped to 12 cryptic exons. Cryptic exons identified in TDP-43–depleted human iPSC–derived neurons were predictive of cryptic exons expressed in postmortem brain tissue from patients with TDP-43 proteinopathy. These cryptic exons produced transcript variants that generated de novo proteins. We found that the inclusion of cryptic peptide sequences in proteins altered their interactions with other proteins, thereby likely altering their function. Last, we showed that 18 de novo peptides across 13 genes were present in CSF samples from patients with ALS/FTD spectrum disorders. The demonstration of cryptic exon translation suggests new mechanisms for ALS/FTD pathophysiology downstream of TDP-43 dysfunction and may provide a potential strategy to assay TDP-43 function in patient CSF.

*Corresponding author. wardme@nih.gov (M.E.W.); p.fratta@ucl.ac.uk (P.F.); petrucelli.leonard@mayo.edu (L.P.).
‡Present address: AstraZeneca, Gaithersburg, MD, USA.
†These authors contributed equally to this work.

**Author contributions:** S.S., Y.A.Q., A.-L.B., L. Petrucelli, P.F., and M.E.W. conceived this study. S.E.K.-H., D.M.R., J.H., J.M.C.-M., V.H.R., M.P.N., J.F.R., and E.K.S. performed human iPSC–derived neuron experiments and other cell culture experiments and developed the TDP-43 knockdown strategy and related assays. A.-L.B., M.Z., M.H., and M.S. performed RNA-seq analyses. S.S., Y.A.Q., C. Bereda, J.R., N.P.S., H.Y., and E.K.S. processed samples and performed and analyzed the human iPSC–derived neuron mass spectrometry experiments. O.G.W. performed and analyzed the Ribo-seq experiments. S.S., A.-L.B., S.I.S., and M.A.N. performed the splicing analyses. S.S., Y.A.Q., C. Bereda, L. Ping, and D.M.D. conducted the CSF proteomics experiments and established the targeted proteomics strategy. A.S. and A.O. provided support in the analysis of the CSF proteomic data. H.Y. provided support for iPSC-derived neuron culture, CSF sample curation, and proteomics experiments. C. Bereda, J.R., Y.-J.Z. performed the APMS experiments. Y.A.Q., S.S., and Z.L. analyzed the APMS data. Y.-J.Z. expressed and purified the MYO18A and HDGFL2 antibodies. Y.-J.Z., S.P., M.J.K., and M.Z. performed and quantified the Western blot experiments. P.R.M. conducted and analyzed the TDP-43 rescue experiments. M.P. performed qRT-PCR experiments. C. Belair, M.M., and A.K. developed the long-read sequencing strategy. M.P., J.Y.K., L. Ping, D.M.D., A.A., A.M., J.D.R., S.J., D.S.R., L.S., D.W.D., J.D.G., N.T.S., and L. Petrucelli coordinated and provided CSF samples and study design and interpretation involving clinical specimens. M.A.N. and L.S. provided data analysis and curation support. S.S., Y.A.Q., A.-L.B., O.G.W., C. Belair, S.I.S., M.A.N., M.P., and M.E.W. visualized the data. Y.A.Q., M.P., H.Y., M.A.N., J.Y.K., S.J., D.S.R., D.W.D., J.U., M.S., J.D.G., A.O., N.T.S., M.M., L. Petrucelli, P.F., and M.E.W. supervised this work. S.S., Y.A.Q., A.-L.B., P.F., L. Petrucelli, and M.E.W. drafted the manuscript. All authors reviewed and edited the manuscript.

**Competing interests:** The participation of M.A.N., Z.L., and S.I.S. in this project was part of a competitive contract awarded to Data Tecnica International LLC by the NIH to support open science research. M.A.N. also currently serves as an advisor for Character Biosciences and Neuron23 Inc. N.P.S. consults for Emtherapro. B.O. serves as a consultant for Columbia University/Tsumura Inc., MediciNova, Biogen, UniQure, Amylyx, and Mitsubishi. L. Petrucelli consults for Expansion Therapeutics.

## INTRODUCTION

Cytoplasmic inclusions of the TAR DNA binding protein 43 (TDP-43) occur in the brains and spinal cords of approximately 97% of amyotrophic lateral sclerosis (ALS), 45% of frontotemporal dementia (FTD), and 40% of Alzheimer's disease (AD) cases (1, 2). Mutations in *TARDBP*, the gene encoding TDP-43, cause familial forms of FTD and ALS, further supporting a central role of TDP-43 in disease pathogenesis (3, 4). TDP-43 mislocalization involves both its clearance from the nucleus and the formation of cytosolic aggregates (1, 2, 5), and both of these events appear to play causal roles in disease pathogenesis. Nuclear clearance of TDP-43 can occur before symptom onset and precedes the formation of cytosolic aggregates (6–8), suggesting that loss of normal TDP-43 function is an early disease mechanism.

TDP-43 has two RNA recognition motifs and directly regulates RNA metabolism by acting as a potent splicing repressor. When TDP-43 splicing repression is lost, the erroneous inclusion of intronic sequences called cryptic exons (CEs) occurs (9). Because CEs are nonconserved intronic sequences, they often introduce frameshifts, premature stop codons, or premature polyadenylation sequences. Such splicing errors can reduce the expression of affected transcripts via nonsense-mediated decay (NMD) or other RNA degradation pathways (10, 11). In turn, the expression of proteins derived from these transcripts often declines in parallel. One such example is the well-characterized CE in *stathmin-2* (*STMN2*), which causes loss of STMN2 mRNA and protein in the cortex and spinal cord of patients with ALS or with frontotemporal lobar degeneration (FTLD; a subset of FTD) (12–14). We and others recently found that TDP-43 also normally represses a CE in *UNC13A*, a critical synaptic gene (15, 16). ALS-linked and FTD-linked risk variants in *UNC13A* reduce the affinity of TDP-43 for *UNC13A* transcripts, thereby promoting CE expression and UNC13A loss in neurons with TDP-43 insufficiency (15). These two examples demonstrate that mis-splicing due to TDP-43 mislocalization can reduce the expression of critical downstream genes, likely affecting neuronal biology during disease.

TDP-43 loss causes widespread CE expression and other related pathological splicing events that affect hundreds of transcripts (9, 13, 14). Certain genes exhibit reduced expression after CE inclusion, but it is also possible that de novo proteins could be synthesized from CE transcripts. Expression of de novo proteins downstream of TDP-43 loss of function could have several important ramifications. Translation of CEs could alter the functions of proteins or induce toxic gain of function with pathophysiological implications. Alternatively, de novo proteins could trigger an autoimmune response, akin to previous observations in autoimmune encephalitis and cancer (17, 18).

Here, we systematically address whether the loss of the TDP-43 function leads to the expression of CE-encoded de novo polypeptides. Using a TDP-43–depleted human induced pluripotent stem cell (iPSC)–derived neuron model of nuclear TDP-43 loss of function, we developed a neuronal CE atlas. We cross-referenced these CEs against a ribosome sequencing (Ribo-seq) dataset from TDP-43–depleted human iPSC–derived neurons and observed extensive ribosome-CE interactions that suggested their active translation. We then developed an unbiased proteogenomic pipeline, using coordinated short-read and long-read

transcriptomic and proteomics analyses of TDP-43–depleted human iPSC–derived neurons, that identified 65 peptides that mapped to 12 CE coding frames. We identified these CEs in postmortem cortex samples from patients with ALS/FTD, validating the physiological relevance of our in vitro cellular models. Through Western blot and proteomic strategies, we confirmed the presence of de novo peptide expression in TDP-43–depleted human iPSC–derived neurons and showed that they can alter the biology of proteins in which they are expressed. Last, we developed a targeted proteomics assay to identify 18 de novo “cryptic” peptides across 13 distinct proteins in cerebrospinal fluid (CSF) samples from patients with ALS/FTD.

## RESULTS

### Ribosomes bind to intronic regions of mis-spliced transcripts in TDP-43–deficient human iPSC–derived neurons

Ribo-seq provides a map of the position and density of ribosomes on individual mRNAs and a proxy for protein translation. If TDP-43–deficient human iPSC–derived neurons translate CEs into de novo proteins (Fig. 1A), then we reasoned that Ribo-seq may reveal ribosome-bound CEs. Using short-read total RNA sequencing (RNA-seq) in TDP-43–depleted human iPSC–derived glutamatergic neurons (fig. S1, A and B, and data file S1), we identified 340 cryptic splicing events across 233 genes out of a total of 498 alternative splicing events (Fig. 1B, fig. S1C, and table S1A). These splicing events included previously identified targets of TDP-43–related splicing repression, such as CEs in *STMN2* and *UNC13A* (13, 15). The majority (70%) of cryptic splicing events were cassette exons, with the remainder comprising exon extension (12%), intron retention (9%), and exon skipping (9%) (fig. S1, C and D, table S1A, and data file S3).

We then queried Ribo-seq data from TDP-43–depleted human iPSC–derived neurons (15) for evidence of changes in ribosome binding within all intronic regions upon TDP-43 knockdown. We identified 30 genes featuring introns with significantly increased footprints in the TDP-43 knockdown Ribo-seq dataset ($P_{adj} < 0.1$; table S1B). When we cross-referenced these sequences with cryptic splicing events identified in Fig. 1B, we found that more than half of these were CE-containing genes (Fig. 1C). By contrast, only five genes had introns with significant decreases in ribosome footprints ($P_{adj} < 0.1$), and none was annotated as containing CEs (table S1B). Next, we selected CEs predicted to generate in-frame peptides and investigated the periodicity of their ribosome footprints. These showed periodicity that was similar to what we observed in annotated exons, suggesting that the signal in CEs was derived from translating ribosomes (Fig. 1D). Overall, TDP-43–depleted human iPSC–derived neurons had roughly nine times more ribosome footprints within CEs than did control human iPSC–derived neurons with normal amounts of TDP-43 (Fig. 1E and data file S1).

Using total RNA-seq and proteomics, we characterized the impact of neuronal TDP-43 loss on CE gene expression at the transcript and protein level (Fig. 1, F to H; fig. S1, E to G; and table S1, C and D). Consistent with prior observations that CEs can reduce mRNA stability (12–16), 37% of CE gene transcripts were reduced in TDP-43–depleted human iPSC–derived neurons (Fig. 1F). We found evidence of additional instability of mis-spliced

genes at the protein level, where 80% of proteins from CE genes displayed decreased expression (Fig. 1G). Analysis of the transcript and peptide abundance of CE-harboring genes at the same time point revealed that whereas some genes (e.g., *SYT7* and *KALRN*) contained out-of-frame CEs and had unstable transcript and protein products, others (e.g., *DNM1* and *MYO18A*) contained in-frame CEs and remained highly expressed (Fig. 1H). Collectively, these data show that whereas CEs often reduce the expression of CE genes, certain CEs and other abnormally spliced transcripts do not affect expression and might be translated into de novo proteins.

## TDP-43–depleted human iPSC–derived neurons express cryptic exons that generate de novo proteins

To identify potential CE-associated de novo proteins at a large scale, we developed an unbiased proteogenomic pipeline that combined proteomics and transcriptomics to identify new peptides (Fig. 2A). Using RNA-seq data from TDP-43–depleted human iPSC–derived neurons, the pipeline returned in-frame amino acid sequences from mis-spliced junctions, henceforth referred to as cryptic peptides. We then added these in silico translated protein sequences of CEs to a standard human proteome reference (data file S2), allowing us to search for trypsin-digested cryptic peptides in shotgun proteomic datasets from TDP-43–depleted human iPSC–derived neurons. This pipeline identified 65 putative trypsin-digested cryptic peptides across 12 genes in TDP-43–depleted human iPSC–derived neurons (Fig. 2B and table S2, A and B), consistent with the possibility that certain CEs are translated into proteins.

More than half of the identified cryptic peptides were predicted to be in frame—and therefore would not cause a premature stop codon—when mapped to short-read RNA-seq of pathologically spliced transcripts (Fig. 2B). One such example is a 46–amino acid, in-frame cryptic peptide in hepatoma-derived growth factor-related protein 2 (HDGFL2) that mapped to a CE in intron 6 (Fig. 2C and fig. S2A). We detected five distinct trypsin-cleaved peptides for HDGFL2 in TDP-43–depleted human iPSC–derived neurons (Fig. 2C). Twenty-three trypsin-cleaved cryptic peptides mapped to out-of-frame CEs in *RSF1*, *KALRN*, and *SYT7* (table S2A). On the basis of the frame of these out-of-frame cryptic peptides, the parent transcripts were predicted to harbor a premature stop codon in the downstream exon. Two additional trypsin-cleaved cryptic peptides mapped to retained introns, whereas four were due to exon skipping (Fig. 2B).

Short-read RNA-seq relies on informatic approaches to stitch together individual reads into full-length transcripts. Errors in predicted full-length sequences can arise when multiple transcript isoforms are coexpressed, particularly in settings of pathological splicing. To unequivocally determine the frame of translation for each cryptic peptide, we performed Nanopore long-read sequencing of control and TDP-43–depleted human iPSC–derived neurons (Fig. 2, D to G; fig. S2, B to D; and table S2C). Mapping all putative cryptic peptides against long-read data from which we could extract the frame, we confirmed four genes (i.e., *HDGFL2*, *AGRN*, *MYO18A*, and *CAMK2B*) with in-frame CEs and two genes (i.e., *MYO1C* and *KCNQ2*) with in-frame exon skipping events in TDP-43–depleted human iPSC–derived neurons (Fig. 2G). An example of an in-frame cryptic peptide from *CAMK2B*

is shown in Fig. 2 (D to F), demonstrating the improved fidelity of long-read sequencing over short-read sequencing in mapping cryptic peptides to complex, pathologically spliced transcripts.

## TDP-43–depleted human iPSC–derived neurons predict mis-splicing in postmortem brain tissue from patients with ALS/FTD

Previously, TDP-43–depleted human iPSC neurons successfully predicted individual CEs that were later validated in patients with ALS/FTD (13–15). Most of the putative cryptic peptides that we observed mapped to CEs that have not been previously described. We therefore tested the fidelity of our TDP-43–depleted human iPSC–derived neuron model in the global prediction of ALS/FTD-associated CEs using three separate approaches. First, we cross-referenced CEs found in TDP-43–depleted human iPSC–derived neurons against a published RNA-seq dataset of fluorescence-activated cell sorting (FACS)–sorted neuronal nuclei from postmortem cortex samples from patients with ALS/FTD (19). Of the 340 cryptic splice junctions that we identified in TDP-43–depleted human iPSC–derived neurons, 230 were detected in FACS-sorted neuronal nuclei from postmortem cortical samples. Some cryptic splice junctions were highly expressed in both TDP-43–positive and TDP-43–negative nuclei, possibly reflecting the cell-type-specific nature of TDP-43–dependent splicing events. Of these, 122 junctions were enriched in nuclei lacking TDP-43. Neuronal nuclei from ALS/FTD postmortem cortex samples clustered according to TDP-43 nuclear presence, rather than by patient demographics (Fig. 3A and table S3, A and B). Using dimensionality reduction analysis of all 230 CEs predicted by TDP-43–depleted human iPSC–derived neurons and detected in FACS-sorted postmortem cortical neuronal nuclei, we found that TDP-43–positive and TDP-43–negative nuclei could be separated on the basis of the first principal component, accounting for 47% of the variability in the dataset (Fig. 3B and data file S1). These data indicate that many cryptic splicing events observed in TDP-43–depleted human iPSC–derived neurons also occur in TDP-43–negative cortical neuronal nuclei derived from ALS/FTD patient brain tissue.

Second, we tested whether cryptic splicing predicted by our TDP-43–depleted human iPSC–derived neuron data could identify cases with TDP-43 pathology in postmortem cortical brain samples. We analyzed bulk RNA-seq datasets from the New York Genome Center (NYGC), comprising 168 frontal and temporal cortex samples from 82 non-neurological disease controls, 20 FTLD–non-TDP cases, and 66 FTLD-TDP cases and 304 motor cortex samples from 49 non-neurological disease controls, 11 ALS non-TDP cases, and 244 ALS-TDP cases. Despite the anticipated low expression and degradation of many CEs by NMD, and the limits of bulk RNA-seq in which only a small proportion of cells have TDP-43 pathology, we detected 298 of the 340 pathologically spliced junctions with at least one spliced read across all the samples (Fig. 3C and table S3, A and B). Next, we tested whether CE expression could predict cases with TDP-43 proteinopathy. Many CE junctions identified in TDP-43–depleted human iPSC–derived neurons had positive prediction power in the FTLD (69) and ALS postmortem samples (25) (Fig. 3C and table S3, A and C to E). We then generated meta-scores by combining expression data from individual or multiple cryptic junctions. We created three different meta-scores: one using the expression of all cryptic events, one using only those junctions with a high predictive power [area under the

curve (AUC) > = 0.6], and one with all predictive cryptic junctions excluding the *STMN2* junction. In the FTLD and ALS postmortem samples, *STMN2*, *KCNQ2*, and all predictive cryptic junctions had the highest power to distinguish cases from controls (Fig. 3D and table S3, A, D, and E).

Third, we used quantitative polymerase chain reaction with reverse transcription (RT-qPCR) to assess CE expression in an independent cohort of 89 FTLD-TDP postmortem cortex samples versus 27 healthy control cortex samples focusing on eight transcripts associated with potential cryptic peptide expression (table S3F). We observed substantially higher expression of each of these CEs in patients with FTLD compared with controls (Fig. 3E, fig. S3A, and data file S1). Thus, we validated our TDP-43–depleted human iPSC–derived neuron data in three independent human datasets.

## Expression of cryptic exons alters protein interactomes

To validate the existence of a de novo protein sequence identified by proteogenomics, we assayed HDGFL2, a protein only modestly reduced in expression in TDP-43–depleted human iPSC–derived neurons (table S1D) and predicted to express an in-frame cryptic peptide (Figs. 2C and 4A). A commercially available polyclonal antibody against canonical HDGFL2 detected a single band on Western blot at the predicted molecular weight in control neurons (Fig. 4B). However, this antibody detected an additional higher molecular weight protein in TDP-43–depleted human iPSC–derived neurons and TDP-43–depleted SH-SY5Y cells, matching the predicted molecular weight of HDGFL2 expressing a CE (HDGFL2-CE) (Fig. 4, B and C; fig. S4, A and B; and data files S1 and S4). We developed an antibody specific to the HDGFL2 cryptic peptide (Fig. 4, A and B, and data file S4) and used a MesoScale Discovery (MSD)–based sandwich immunoassay to accurately quantify this cryptic peptide in TDP-43–depleted human iPSC–derived neurons (fig. S4C and data file S1). We used immunofluorescence staining to confirm that the expression of the HDGFL2-CE peptide was restricted to cells with loss of nuclear TDP-43 expression (Fig. 4D). We next used small interfering RNAs (siRNAs) to knock down TDP-43 in human embryonic kidney (HEK) 293 cells harboring a doxycycline-inducible, siRNA-resistant green fluorescent protein (GFP)–TDP-43 transgene (20). Induction of siRNA-resistant GFP–TDP-43 prevented HDGFL2 cryptic-peptide production, confirming that the expression of the cryptic protein was due to TDP-43 loss of function (fig. S4, D and E, table S4A, and data file S1).

We developed an additional antibody against an in-frame cryptic peptide in MYO18A (fig. S5, A and B). Using this antibody in a Western blot assay, we detected a protein of the anticipated molecular weight of MYO18A expressing a CE (MYO18A-CE) in TDP-43–depleted human iPSC–derived neurons but not in control human iPSC–derived neurons or those depleted of fused in sarcoma (FUS), an ALS/FTD-associated RNA binding protein that does not cause loss of TDP-43 function (Fig. 4, E and F; fig. S5, C and D; and data files S1 and S4). These data validate our predictions that some CEs are translated into de novo proteins expressed highly enough to be detected by standard immunoassays and are specific to loss of TDP-43 function.

We reasoned that the inclusion of a 46–amino acid cryptic peptide in HDGFL2—a protein thought to regulate chromatin and DNA repair in non-neuronal cells (21)—might alter its interacting partners and thus its biology. We performed affinity purification mass spectrometry of HDGFL2 in control and TDP-43–depleted human iPSC–derived neurons, using an anti-HDGFL2 antibody to immunoprecipitate HDGFL2 and its associated proteins. We identified 178 HDGFL2-interacting proteins that were significantly enriched in anti-HDGFL2 antibodies compared to control immunoglobulin G (IgG) pull-down assays ($P_{adj} < 0.05$; Fig. 4G and table S4, B and C). Gene Ontology (GO) term analysis of HDGFL2-interacting proteins in neurons revealed strong enrichments in the ribosome, spliceosome, actin, and neurodegeneration-related pathways (fig. S6, A and B). Unexpectedly, MYO18A was a top interacting protein of HDGFL2, clustering with other actin-regulating HDGFL2 interactors (fig. S6C). We then analyzed how TDP-43 loss—which causes expression of HDGFL-CE—altered the HDGFL2 interactome. Sixteen proteins increased their relative interactions with HDGFL2 upon TDP-43 loss, a large fraction of which regulate mRNA splicing, chromatin remodeling, and DNA repair (Fig. 4, G and H). In contrast, 13 proteins decreased their relative interactions with HDGFL2 upon TDP-43 loss. Most of these proteins play roles in cytoskeleton organization, with five comprising core regulators of Arp2/3 actin nucleation (Fig. 4H). We next directly tested whether the interactomes of CE-HDGFL2 and full-length native HDGFL2 (FL-HDGFL2) differed by expressing CE-HDGFL2-myc-flag or FL-HDGFL2-myc-flag in HEK293 cells followed by affinity purification mass spectrometry using an anti-flag antibody. We observed substantial differences in the interactomes of CE-HDGFL2-myc-flag and FL-HDGFL2-myc-flag (fig. S6, D to G, and table S4, D and E). CE-HDGFL2 displayed increased interactions with RNA binding proteins, including a small number of splicing-regulating proteins, consistent with our observation from endogenous HDGFL2 pull-down assays of TDP-43–depleted human iPSC–derived neurons (fig. S6, E and G). As expected, CE-HDGFL2 had reduced interactions with proteins that regulated the actin cytoskeleton (fig. S6, F and H). These observations demonstrate that the inclusion of CE sequences in a translated protein can alter its protein-protein interactome downstream of TDP-43 loss of function.

## Validation of cryptic peptides by targeted proteomics

Because antibody development against new peptides is a time- and resource-intensive endeavor, we developed a targeted mass spectrometry–based approach to validate and measure additional cryptic peptides predicted by proteogenomics. By co-injecting endogenous cryptic peptides and a panel of their stable isotope (SIL) heavy peptide standards coupled with parallel reaction monitoring (PRM) mass spectrometry proteomics, we quantified dozens of peptides in a single sample run (Fig. 5A). We designed PRM assays for the 65 cryptic peptides predicted by our proteogenomic pipeline (table S5A). Of these, we successfully quantified 12 endogenous trypsin-digested cryptic peptides across four genes in TDP-43–depleted human iPSC–derived neurons (Fig. 5, B to D, and fig. S7A) that were precisely co-eluted with their SIL counterparts (Fig. 5B and fig. S7A) and displayed nearly identical fragmentation spectra to the corresponding heavy peptides (Fig. 5C and fig. S7A). The spectral plot and corresponding mass spectra for an *SYT7* cryptic peptide are shown in Fig. 5 (B and C); peptides from *CAMK2B*, *MYO18A*, and *RSF1* could also be detected (fig. S7A). Consistent with CE mRNA expression, we found that

cryptic peptides from all four genes were highly increased in TDP-43–depleted human iPSC–derived neurons, with almost no expression in control human iPSC–derived neurons (Fig. 5E, table S5B, and data file S1). These experiments suggest that multiple genes express cryptic peptides in the setting of functional TDP-43 loss and demonstrate that proteomics can detect and measure the expression of cryptic peptides in complex biological samples.

### Identification of cryptic peptides in human CSF samples

We next asked whether patients with ALS/FTD spectrum disorders associated with TDP-43 mislocalization express cryptic peptides. Because most CE genes encode intracellular proteins, it was unclear whether proteins encoded by these genes were present in clinically relevant biofluids, such as CSF. We compared a shotgun proteomics dataset from ALS CSF samples and our CE dataset from TDP-43–depleted human iPSC–derived neurons (table S6A). Canonical proteins from 47 CE genes were present in ALS CSF samples, including several that expressed cryptic peptides in TDP-43–depleted human iPSC–derived neurons (Fig. 6, A and B, and table S6C). We searched two additional ALS CSF proteomics datasets from independent patient cohorts and identified four additional CE genes with canonical proteins expressed in human CSF (Fig. 6B; fig. S8, A and B; and table S6, B to F). These observations provided a rationale for developing assays that could detect cryptic peptide variants of these proteins in ALS/FTD CSF samples.

We augmented our previous list of SIL peptide standards with these additional candidates and used data-independent acquisition (DIA) proteomics to co-measure 65 endogenous and heavy peptide standards in ALS/FTD patient CSF in a single mass spectrometry run (fig. S8C and table S5A). We successfully detected 18 tryptic peptides across 13 genes that mapped to CEs in the CSF of patients with ALS/FTD spectrum disorders (Fig. 6, C to E; fig. S9A; and table S6, G to I). These endogenous tryptic-cryptic peptides co-eluted with their heavy peptide standards (Fig. 6C and fig. S9A) and had matching mass spectra (Fig. 6D). We monitored the expression of these peptides in CSF from 15 patients with ALS/FTD. Most peptides were unequivocally observed in multiple patients, with 10 peptides across eight genes detectable in >80% of patients (Fig. 6E and table S6, H and I). Cryptic peptides were also detected in healthy control individuals, but variation across mass spectrometry runs and limited sample sizes precluded quantitative comparison between control individuals and patients with ALS/FTD (table S6J). These data indicate that, as in TDP-43–depleted human iPSC–derived neurons, CEs are expressed and translated in the CNS of patients with ALS/FTD.

## DISCUSSION

TDP-43 loss of function causes widespread cryptic splicing that can reduce the expression of mRNA transcripts. Here, we demonstrate that such mis-splicing can also generate de novo proteins from pathological transcripts. Our studies indicate that the inclusion of CEs in proteins can, in certain instances, alter their interacting protein partners. We show that TDP-43–depleted human iPSC–derived neurons can be used to model CEs found in ALS/FTD brains, enabling the prediction of cryptic peptides and the development of antibody and mass spectrometry–based methods for their detection. Last, we demonstrate

the presence of TDP-43–related de novo peptides in CSF samples from patients with ALS/FTD spectrum disorders.

These findings are relevant to ALS/FTD pathophysiology. In addition to the well-described loss-of-function mechanisms downstream of cryptic splicing, new possibilities of altered protein function should now be considered. For example, we found that TDP-43 loss increased HDGFL2 interaction with splicing regulators while decreasing its interaction with other protein networks, such as those that regulate actin dynamics. This suggests both the gain and the loss of function of HDGFL2 occur as a direct consequence of CE expression. Although these observations hint at the possibility that TDP-43 loss of function may alter the biology of functionally inter-related proteins, further studies are needed to address whether CE-harboring proteins play a direct role in disease pathogenesis or are an epiphenomenon. In addition to TDP-43 loss of function, cytoplasmic aggregation of TDP-43 is a known driver of toxic gain of function that can induce cell death and may exacerbate functional TDP-43 loss by modulating its capacity to regulate RNA biology (22–24). Because transient mislocalization and formation of aggregate-like inclusions of TDP-43 occur in settings of cell stress (e.g., after axotomy or oxidative stress), these findings also imply that CE expression downstream of TDP-43 mislocalization could play a regulatory role in settings outside of disease pathophysiology (25–27). We additionally found that HDGFL2 interacts with two other targets of TDP-43–related splicing regulation: MYO18A, a CE-harboring member of the actin regulatory network, and POLDIP3, which exhibits exon skipping in settings of TDP-43 loss. Whether functional relationships exist between genes that form CEs upon TDP-43 mislocalization will need to be explored.

Our data suggest that only a small fraction of CE-harboring transcripts produce stable polypeptides, whereas most undergo degradation at the RNA or protein level. Of 154 CE-harboring genes, only 27 contained in-frame insertions, which are less likely to trigger RNA degradation pathways such as NMD. Additional quality control pathways may regulate the degradation of misfolded proteins translated from CEs. Despite high ribosome occupancy of *STMN2* intronic sequences, no STMN2 de novo peptides were identified in our proteomic studies, suggesting that any proteins produced from the *STMN2* CE are highly unstable or poorly detectable by mass spectrometry.

Our findings also provide a rationale for future evaluation of the potential role of cryptic peptides in autoimmune dysregulation in neurodegenerative disorders. It is conceivable that cryptic peptides could be presented by MHCs for recognition by cytotoxic T cells. Infiltration of cytotoxic $CD8^+$ T cells has been previously reported in the brain and spinal cord of patients with ALS (28–30). The possibility that cryptic peptides could elicit the expression of autoantibodies warrants further consideration. Neo-antigens are now well described for other neurodegenerative disorders, including Lewy body dementia and Parkinson’s disease, and suggest the possibility of searching for T cell or antibody-based responses against such antigens (31, 32). Retrospective analysis of stored serum samples recently led to the seminal discovery of Epstein-Barr virus as the viral etiology of multiple sclerosis (33, 34). Similar approaches may be fruitful for the study of CE-related autoantibodies in ALS.

To accurately map putative cryptic peptides to mRNA transcripts, we generated a transcriptome-wide database of long-read sequences in TDP-43–depleted human iPSC–derived neurons. Whereas informatics methods to analyze long-read data for splicing abnormalities are still in development, we anticipate that these datasets will be of immediate use to the community. Peptide libraries can be generated from our proteomic datasets and used to probe for autoreactive B or T cell populations and antibodies in patient biospecimens. The expanded identification of CE-harboring transcripts within patient postmortem brain tissue may prove useful in understanding the pathophysiology of ALS/FTD and could be leveraged in the development of mRNA-based diagnostics. We provide validated amino acid sequences for cryptic peptides, which could serve as the basis for designing high-sensitivity antibody-based or mass spectrometry–based assays [e.g., SIMOA (single molecular array), SRM (selected reaction monitoring), and IP-MS (immunoprecipitation followed by mass spectrometry)].

Currently, there are no methods that enable the identification of living patients with TDP-43 pathology nor any means to monitor responses to TDP-43–directed therapies that are in development for ALS/FTD spectrum disorders. Approximately 20 to 50% of patients with AD—and 75% of patients with severe AD—also have TDP-43 co-pathology and may represent an additional patient subpopulation that could benefit from drugs targeting pathogenic TDP-43 (35–37). Identification of imaging-based techniques to detect TDP-43 or its pathological forms has proven challenging. Antibody-based strategies to monitor the abundance and biochemical properties (e.g., phosphorylation and solubility) of TDP-43 have been attempted, but the interpretability and theragnostic utility of these assays in clinical settings are uncertain. A critical unmet need to develop molecular readouts of TDP-43 pathology therefore exists, both to guide disease stratification and to enable better deployment of TDP-43–directed therapies.

Given that mis-splicing is a direct consequence of TDP-43 dysfunction, assays that monitor cryptic peptides in biospecimens could be useful for monitoring therapeutic responses. Targeted mass spectrometry–based approaches could be optimal for developing ultrasensitive bioassays for clinical application. An example of this is SRM, wherein synthetic stable isotope-labeled peptides are used as internal standards to enable the quantification of trace analytes (38). Alternate strategies may leverage high-sensitivity antibodies or aptamers against native proteins expressing cryptic peptides. Such tools could subsequently enable SIMOA assays that can measure analytes at the single-molecule level (39). New methods to enrich brain-derived exosomes in plasma and paired single-molecule detection strategies (e.g., SMAC) are being developed and open exciting possibilities for noninvasive disease profiling in the blood (40). Ultimately, a multi-analyte panel for cryptic peptides may be an effective diagnostic strategy. This is consistent with our findings of improved disease predictability in postmortem brain tissue upon consideration of multiple CEs over any single mis-spliced junction (except for *STMN2* CE, which cannot yet be detected at the protein level).

Our study has several limitations. We were unable to conclusively determine whether cryptic peptides are expressed more highly in cases compared with healthy controls and therefore cannot conclude that these peptides are specific for diseases involving TDP-43 pathology.

Our mass spectrometry assays were designed to enable specific detection of target peptides; however, in their current form, they are only semiquantitative and subject to technical artifacts when analyzing the absolute abundance of peptides in biofluid samples with high dynamic ranges and low peptide abundances. Future development of more sensitive, quantitative proteomic, and antibody-based assays will enable reliable comparisons of cryptic peptide abundances across patients with ALS/FTD and non-disease controls, as well as across different patient populations, such as those with AD, dementia with Lewy bodies, and Parkinson's disease with coexistent TDP-43 pathology. Our assays also require the depletion of the top 14 contaminants in CSF, which introduces additional technical variability to sample preparation. SRM assays are the gold standard for absolute protein quantification and enable target peptides to be measured without immunodepletion, but the development and optimization of such assays are laborious and specialized. The development of specific cryptic peptide assays is further challenged by the current lack of a definitive antemortem TDP-43 assay. Perceived "healthy" controls could harbor early TDP-43 pathology. Approximately 20 to 50% of all AD cases exhibit concomitant TDP-43 pathology before the onset of overt symptoms (35, 36). In addition, there are no clearly defined antemortem pathological markers of limbic predominant age-related TDP-43 encephalopathy (LATE). These challenges underscore a need for careful selection of control populations. Genetic markers, such as *C9orf72* repeat expansion, can be used to predict the presence of TDP-43 aggregates or suggest their absence in the setting of rare genetic markers, such as *FUS* and *SOD1* mutations. As a pure tauopathy, progressive supranuclear palsy may represent a superior control population to age-matched "healthy" individuals. With the current datasets at hand, our hope is that multiple efforts to develop either mass spectrometry– or antibody-based methods for absolute quantification of cryptic peptides can ensue in parallel.

In summary, we show de novo cryptic peptides occurring downstream of TDP-43 dysfunction in cellular models and ALS/FTD patient CSF samples. These findings should guide new discoveries on the role of cryptic peptides in diseases involving TDP-43 pathology and provide a framework for the development of clinical-grade tests to measure TDP-43 function in biospecimens.

## MATERIALS AND METHODS

### Study design

The objective of this study was to determine whether CEs generate de novo peptides in the setting of TDP-43 deficiency. To develop a neuronal CE catalog, we performed total RNA-seq and differential splicing analyses on human iPSC–derived neurons with TDP-43 depleted by CRISPRi. We used these cells to examine transcriptional and splicing changes due to loss of nuclear TDP-43 function. We then designed an unbiased proteogenomic pipeline, using parallel RNA-seq and shotgun proteomics in the TDP-43–depleted human iPSC–derived neurons, to identify peptides that map to CE coordinates. Next, we cross-referenced these predicted CEs against datasets from ALS/FTD postmortem brain tissue to test the fidelity of our in vitro model in predicting TDP-43 pathology in human brains. We used Western blot and targeted proteomics to validate the presence of de novo peptides in

TDP-43–depleted human iPSC–derived neurons. We then examined the impact of de novo peptides on the biology of the proteins in which they were expressed via affinity purification mass spectrometry. Last, we developed a targeted proteomics assay to demonstrate the presence of de novo “cryptic” peptides in CSF samples from patients with ALS/FTD spectrum disorders. The sample size was determined empirically in accordance with field standards to ensure sufficient power for detecting statistical differences. The number of experimental replicates is provided in each figure legend. Only RNA and proteins of low abundance or quality were excluded. All ALS/FTD CSF samples that were provided to us by the Mayo Clinic were analyzed, and none was excluded. Written informed consent was obtained from all participants or their family members to collect and analyze CSF samples.

### CRISPRi knockdown experiments

CRISPRi knockdown experiments were performed in the WTC11 human iPSC line harboring stable TO-NGN2 and dCas9-BFP-KRAB cassettes at safe harbor loci (41). CRISPRi knockdown of TDP-43 was achieved by infecting iPSCs with a lentiviral-expressed single guide RNA (sgRNA) targeting the transcription start site of TARDBP (or nontargeting control sgRNA) (15). Neurons were differentiated from TDP-43–depleted human iPSCs as described previously (42, 43) and harvested 17 days after differentiation for all experiments.

### Short-read RNA sequencing, differential gene expression, and splicing analysis

RNA was extracted from day 17 TDP-43–depleted human iPSCs using the Direct-zol RNA Miniprep Kit (Zymo Research R2051) and sequenced on a NovaSeq 6000 (2 × 150–base pair paired-end). Sequencing files were trimmed for adapters (cutadapt v2.5) (44), quality checked (FastQC v0.11.6) (45), and aligned to GRCh38 reference genome (STAR v2.7.3a) (46). Differential expression analysis was performed using the standard DESeq2 workflow (47). For differential splicing analysis, all samples were run through MAJIQ (48) in the same manner using a custom Snakemake pipeline (https://github.com/frattalab/splicing). An additional pipeline was developed to visualize and categorize each mis-spliced junction as CE, exon skipping, intron retention, or canonical junction (https://github.com/NIH-CARD/proteogenomic-pipeline).

### Long-read sequencing and data processing

Sequence-specific and regular cDNA-PCR libraries were prepared using the Oxford Nanopore Technologies sequencing kit (SQK-PCS109) and sequenced on a PromethION device. Sequencing data were basecalled (Guppy v3.4.5) and mapped to GRCh38 reference genome (minimap2) (49). Basecalled reads were also independently aligned against the human transcriptome (Ensembl version 92). Transcript identification and quantification were performed with Bambu at maximum sensitivity (50). Only reads overlapping mis-spliced genes were used.

### Ribosome profiling

Ribosome profiling libraries from three biological replicates both of control and TDP-43–depleted human iPSC–derived neurons were obtained from a previous study (15) but

underwent Cas9-based rRNA depletion using Ribocutter (51), followed by resequencing on an Illumina Hi-Seq 4000 machine to improve read depth. Multi-mapping reads were discarded, and reads 28 to 30 nucleotides in length were selected for analysis.

### De novo peptide sequence prediction

Dasper was used to classify each mis-spliced junction as a novel acceptor, novel donor, or exon skipping event (52). Exonic regions were identified using Gencode v31 reference annotation and two different transcript assembly tools, Scallop (53), and Stringtie2 (54). All novel exonic regions were mapped back to coordinates that overlapped cryptic junctions. Transcripts overlapping with cryptic events were identified as “backbone transcripts” and used to insert the novel cryptic event. The resulting genomic regions were used to extract nucleotide sequences. For each nucleotide sequence, the amino acid sequence of all possible open reading frames (ORFs) between every methionine and stop codon was extracted.

### Protein structure prediction

To assess the potential impact of the inclusion of an additional coding exon to HDGFL2 (Q7Z4V5) and MYO18A (Q92614) protein structures, we exploited Alphafold2 (10.1038/s41586-021-03819-2) using the monomer settings. The input sequences for the two proteins were computed by adding the cryptic peptide sequences to the canonical sequences from UniProt. The resulting structure files were visualized on https://molstar.org/viewer/.

### Liquid chromatography and mass spectrometry

Protein tryptic digestion was performed using a fully automated sample preparation workflow as previously described (55). For human CSF samples, the top 14 most abundant proteins were depleted with a High Select depletion spin column and resin (Thermo Fisher Scientific, A36369). Three data acquisition approaches (DDA, DIA, and PRM) were used in liquid chromatography–tandem mass spectrometry (LC-MS/MS) analyses. TDP-43–depleted human iPSC–derived neuron samples were separated on a Waters AQUITY UPLC M_Class system and injected into an Orbitrap Exploris 480 for DIA proteomics. All other samples were separated on an UltiMate 3000 RSLCnano system and injected into a high-resolution Orbitrap Eclipse MS. Protein tryptic digestion was performed using a fully automated sample preparation workflow as previously described (55). For human CSF samples, the top 14 most abundant proteins were depleted with a High Select depletion spin column and resin (Thermo Fisher Scientific, A36369). Three data acquisition approaches (DDA, DIA, and PRM) were used in LC-MS/MS analyses. See Supplementary Materials and Methods for additional details.

### Proteomics database search and statistical analysis

DDA-based discovery proteomics was performed using a custom database of de novo peptides (data file S2) in PEAKS studio v10.6 (56). Skyline software (57) was used for the quantification and visualization of PRM and DIA targeted proteomics data. Statistical analyses were conducted in R studio, using a two-sided *t* test and Benjamini-Hochberg adjustment for multiple comparisons. Differential expression analysis was carried out in

the Spectronaut software (58), generating log fold changes and *q* values for volcano plot analysis. See Supplementary Materials and Methods for additional details.

### Splicing analysis of postmortem brain tissue

Data from FACS-sorted frontal cortex neuronal nuclei were obtained from the Gene Expression Omnibus database (GSE126543) and aligned to the GRCh38 reference genome as previously described (15). The splice junction output tables were then clustered and converted into PSI metrics using Dasper (52). Because splicing tools can be prone to one-off errors for exact splice junction coordinates, the 340 bona fide splicing events from MAJIQ were manually curated against the automated splice junctions from STAR to confirm the absence or presence of each event in the FACS-sorted nuclei.

The analysis of postmortem brain tissue from NYGC contained 472 neurological tissue samples from 286 individuals in the NYGC ALS dataset, including non-neurological disease controls, FTLD, ALS, FTD with ALS (ALS-FTLD), and ALS with suspected AD (ALS-AD). The NYGC dataset was analyzed for the 340 bona fide splicing events from MAJIQ using a modified splice junction parsing pipeline. Junctions from NYGC samples that overlapped with the 340 events were manually curated to ensure accuracy. Read counts for each splice junction were normalized and converted to *z*-scores to calculate AUC scores for TDP-43 pathology classification performance using the pROC package (59). Meta-scores were created using *z*-scores across all cryptic junctions, junctions with a positive predictive value above 0.6, and predictive junctions excluding STMN2. See Supplementary Materials and Methods for additional details.

### qRT-PCR validation of cryptic exons

Quantitative real-time PCR (qRT-PCR) was conducted using SYBR GreenER qPCR SuperMix (Invitrogen) for all samples in triplicates. qRT-PCR was run in a QuantStudio 7 Flex Real-Time PCR system (Applied Biosystems). Relative quantification was determined using the ΔΔCt method and normalized to the endogenous controls *GAPDH* and *RPLP0*. Primer efficiency was verified for each CE before running the qRT-PCRs. See table S3E for a list of primers. Additional details are provided in Supplementary Materials and Methods.

### Western blot

MYO18A-CE antibody was generated by LabCorp by immunizing rabbits with a peptide including the complete 20-residue neoepitope (VKEEDKTLPKPGSPGKEEGA). Equal amounts of protein were loaded into 10-well 3 to –8% tris-acetate (MYO18A and MYO18A-CE) or 10-well 10% tris-glycine gels (HDGFL2, HDGFL2-CE, TDP-43, FUS, and GAPDH) (Thermo Fisher Scientific) and transferred to membranes. The primary antibodies used for the MYO18A-CE experiment were anti-rabbit MYO18A-CE antibody (1:500), anti-rabbit TDP-43 antibody (1:1000; Proteintech, 12892-1-AP), and anti-mouse GAPDH antibody (1:5000; Meridian Life Science, H86504M). For the quantification of total HDGFL2 protein, equal amounts of protein were loaded in 7% bis-tris gels and run with Mops buffer (Thermo Fisher Scientific) and transferred to membranes (Amersham GE Healthcare). Primary antibodies used were rabbit anti-HDGFL2 (1:1000; Sigma-Aldrich, HPA 044208), mouse anti-TDP-43 (1:5000; Abcam, ab104223), and rat anti-tubulin

(1:5000; Millipore, MAB1864). For FUS studies, an anti-mouse FUS antibody (1:500; Santa Cruz Biotechnology, SC-47711) was used. Western blots were developed on a Bio-Rad ChemiDoc and quantified with ImageJ. See Supplementary Materials and Methods for additional details.

### MSD assay

HDGFL2 cryptic protein quantification in TDP-43–depleted human iPSC–derived neuron lysates was carried out using an MSD-based sandwich immunoassay. Briefly, TDP-43–depleted human iPSC–derived neuron lysates were diluted in TBS. Serial dilutions of TDP-43–depleted human iPSC–derived neuron lysates (0, 0.5, 1, 2, 4, and 8 μg) were tested in duplicate wells. The wild-type HDGFL2 antibody at a concentration of 4 μg/ml (Proteintech, #15134-1-AP) was used as the capture antibody. The Sulfo-tagged–HDGFL2-CE antibody at a concentration of 4 μg/ml was used as the detection antibody. Response values corresponding to the intensity of emitted light on electrochemical stimulation of the assay plate using the MSD QUICKPLEX SQ120 were acquired.

### TDP-43 GFP rescue experiments

GFP TDP-43 inducible cell lines were previously generated as previously described (20). Briefly, Flp-In HEK293 was transfected with siRNA against TARDBP for 72 hours. In the rescue experiments, 24 hours before the collection of cells and protein lysates, the medium was replaced with Dulbecco's modified Eagle's medium with doxycycline to induce the siRNA-resistant N-terminally GFP-tagged TDP-43.

### Immunofluorescence staining of HDGFL2 cryptic peptide in TDP-43–depleted human iPSC–derived neurons

Three days after differentiation, TDP-43–depleted human iPSC–derived neurons were dissociated using Accutase and then seeded on poly-L-ornithine–coated glass coverslips in a 24-well plate at a density of $2 \times 10^4$ cells per well in maturation medium. Fourteen days later, the neurons were fixed with 4% paraformaldehyde. The fixed neurons were permeabilized with 0.5% Triton X-100 for 10 min, blocked with 10% normal goat serum in phosphate-buffered saline (PBS) for 1 hour, then incubated with the anti-rabbit HDGFL2-CE antibody (1:500) and anti-mouse TDP-43 antibody (1:500) overnight at 4°C. After washing, sections or cells were incubated with corresponding Alexa Fluor 488–conjugated donkey anti-rabbit antibody (1:500) and Alexa Fluor 568–conjugated donkey anti-mouse antibody (1:500) for 2 hours. One percent normal goat serum in PBS was used to dilute the primary and secondary antibodies. Hoechst 33258 was used to stain cellular nuclei. Images were obtained on a Zeiss LSM 980 laser scanning confocal microscope.

### Affinity purification mass spectrometry

Anti-IgG or anti–HDGFL2-CE antibodies were added to cell lysates and incubated overnight. Sample protein concentrations were evaluated using a detergent compatible protein assay (Bio-Rad, Hercules, CA, catalog no. 5000111). Automated affinity purification mass spectrometry was performed using an automated protocol on the KingFisher Flex purification system, followed by overnight incubation. Peptides were dried and reconstituted

in a 2% acetonitrile and 0.4% trifluoroacetic acid solution. Peptide concentrations were evaluated on a Denovix DS-11 FX spectrophotometer/fluorometer and normalized for analysis by LC-MS/MS. Precursor matching, protein inference, and quantification were performed in the Spectronaut software (58) using the DirectDIA workflow in default settings. Differential abundance analysis was carried out in the Spectronaut software 16.2 (58), generating log fold changes and *q* values by using a two-sided *t* test and Benjamini-Hochberg adjustment for multiple comparisons. R software version 4.2.3 was used for data visualization, and ShinyGo was used for GO analysis. See Supplementary Materials and Methods for additional details.

### STRING analysis

GO analysis was performed on significantly enriched HDGFL2-interacting proteins ($P_{adj} < 0.05$), compared with control IgG pulldowns, or HDGFL2-interacting proteins that were significantly changed in their interactions depending on TDP-43 knockdown ($P_{adj} < 0.05$). GO term analysis was performed in ShinyGo v0.76.3 (60). String analysis was performed using String v11.5 (61), and resulting networks were downloaded and visualized using Cytoscape (v3.9.1) (62).

### CSF samples

CSF samples were collected through the Neurological Disease Biorepository and Biomarker Initiative at the Mayo Clinic's campus in Jacksonville, Florida (IRB #13-004314 and IRB# 11-002986). Our cohort of cases consisted of nine males and six females (Caucasian, and not of Hispanic ethnicity), with a median age at CSF collection of 64 years (range: 50 to 78) and a median age of ALS onset of 59 years (range: 49 to 79 years).

### Statistical analysis

Graphs were generated using GraphPad Prism version 9.3.1. Data are displayed as means ± SEM; individual data points are also shown. Statistical testing was performed using GraphPad Prism version 9.3.1 and R statistical programming language. Shapiro-Wilk test was used to assess normality. *P* values for normally distributed datasets were determined by unpaired *t* test, with Benjamini-Hochberg adjustment for multiple comparisons. For nonnormally distributed data, nonparametric analysis was carried out using the Mann-Whitney test or one-way analysis of variance (ANOVA). $P < 0.05$ was considered significant for two-tailed tests, whereas $P < 0.1$ was considered significant for one-tailed tests. The number of biological replicates and details of statistical analyses are provided in the figure legends.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Authors

Sahba Seddighi[1,2,†], Yue A. Qi[3,†], Anna-Leigh Brown[4,†], Oscar G. Wilkins[4,5], Colleen Bereda[3], Cedric Belair[6], Yong-Jie Zhang[7,8], Mercedes Prudencio[7,8],

Matthew J. Keuss[4], Aditya Khandeshi[6], Sarah Pickles[7,8], Sarah E. Kargbo-Hill[1], James Hawrot[1], Daniel M. Ramos[3], Hebao Yuan[1], Jessica Roberts[3], Erika Kelmer Sacramento[9], Syed I. Shah[10], Mike A. Nalls[3,10], Jennifer M. Colón-Mercado[1,‡], Joel F. Reyes[1], Veronica H. Ryan[1], Matthew P. Nelson[3], Casey N. Cook[7,8], Ziyi Li[3,10], Laurel Screven[3], Justin Y. Kwan[1], Puja R. Mehta[4], Matteo Zanovello[4], Martina Hallegger[5,11], Anantharaman Shantaraman[12], Lingyan Ping[12], Yuka Koike[7,8], Björn Oskarsson[7,8], Nathan P. Staff[13], Duc M. Duong[12], Aisha Ahmed[4], Maria Secrier[14], Jernej Ule[5,11], Steven Jacobson[1], Daniel S. Reich[1], Jonathan D. Rohrer[4], Andrea Malaspina[4], Dennis W. Dickson[7,8], Jonathan D. Glass[15], Alessandro Ori[9], Nicholas T. Seyfried[12,16], Manolis Maragkakis[6], Leonard Petrucelli[7,8,*], Pietro Fratta[4,5,*], Michael E. Ward[1,3,*]

## Affiliations

[1]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA.

[2]Nuffield Department of Population Health, University of Oxford, Oxford, UK.

[3]Center for Alzheimer's and Related Dementias (CARD), National Institute on Aging and National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA.

[4]UCL Queen Square Motor Neuron Disease Centre, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, UCL, London, UK.

[5]Francis Crick Institute, London, UK.

[6]Laboratory of Genetics and Genomics, National Institute on Aging, Intramural Research Program, National Institutes of Health, Baltimore, MD, USA.

[7]Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA.

[8]Neuroscience Graduate Program, Mayo Clinic Graduate School of Biomedical Sciences, Jacksonville, FL, USA.

[9]Leibniz Institute on Aging, Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena, Germany.

[10]Data Tecnica International, Washington, DC, USA.

[11]UK Dementia Research Institute at King's College London, London, UK.

[12]Department of Biochemistry, Emory University School of Medicine, Atlanta, GA, USA.

[13]Department of Neurology, Mayo Clinic, Rochester, MN, USA.

[14]Department of Genetics, Evolution and Environment, UCL Genetics Institute, UCL, London, UK.

[15]Department of Neurology, Center for Neurodegenerative Diseases, Emory University, Atlanta, GA, USA.

[16]Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA.

## Acknowledgments:

We thank A. Singleton, M. Cookson, and C. Pantazis for helpful input. We thank M. Guo for assistance with CSF experiments and gratefully acknowledge support from the Friedrich Loeffler Institute (FLI) Core Facility Proteomics.

**Funding:**

This work was supported, in part, by the Intramural Research Program of the National Institutes of Neurological Disorders and Stroke (to M.E.W. and S.I.S.); by the Center for Alzheimer's and Related Dementias, National Institute on Aging, and National Institute of Neurological Disorders and Stroke (to M.E.W.); by the Robert Packard Center for ALS Research (to M.E.W.); by the Chan Zuckerberg Initiative (to M.E.W. and A.O.); by Target ALS (to M.E.W., P.F., L. Petrucelli, and M.P.); by Muscular Dystrophy Association and National Institute on Aging (to J.D.G.); by National Institutes of Neurological Disorders and Stroke (to J.D.G. and N.T.S.); and by NIH grant no. T32 GM136577 (to S.S.). S.S. is additionally supported by the NIH Oxford-Cambridge Scholars Program. P.F. is supported by a UK Medical Research Council Senior Clinical Fellowship and the MNDA Lady Edith Wolfson Fellowship (MR/M008606/1 and MR/S006508/1), NIH grant # U54NS123743, Target ALS, and the Robert Packard Center for ALS Research. The FLI is a member of the Leibniz Association and is financially supported by the Federal Government of Germany and the State of Thuringia. This work was also supported by Mayo Clinic Foundation (to L. Petrucelli), the Liston Family Foundation (to L. Petrucelli), NIH/National Institute on Aging grants ADRC 5P30AG0062677 (to L. Petrucelli), ALLFTD U19AG063911 (to L. Petrucelli), R01AG062171 (to L. Petrucelli), and RF1AG062077 (to L. Petrucelli); NIH/National Institute of Neurological Disorders and Stroke grants U54NS123743 (to L. Petrucelli), R35NS097273 (to L. Petrucelli), and P01NS084974 (to L. Petrucelli); Alzheimer's Association Zenith Award (to L. Petrucelli); and Cure Alzheimer's Fund (to L. Petrucelli). V.H.R. was supported by NIH grant no. FI2GM142475. P.R.M. is supported by a Wellcome Trust Clinical Training Fellowship (102186/B/13/Z). This work was also supported by the UK Dementia Research Institute award #UK DRI-RE13553, which receives its funding from UK DRI Ltd., funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK, and by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC0102), the UK Medical Research Council (CC0102), and the Wellcome Trust (CC0102).

## Data and materials availability:

HDGLF2 and MYO18A-CE antibodies can be obtained from L. Petrucelli through a material transfer agreement. All pipelines and codes used are deposited on Zenodo (DOI 10.5281/zenodo.10479046). RNA-seq, Ribo-seq, and proteomics data files can be accessed at Alzheimer's Disease Workbench (ADWB): https://fair.addi.ad-datainitiative.org/#/data/datasets/mis_spliced_transcripts_generate_de_novo_proteins_in_tdp_43_related_als_ftd_00005. ALS CSF 2D-LC-MS/MS data and ALS CSF TMT-MS data can be downloaded from Synapse (SynID: syn26642919 and syn25795030, respectively): www.synapse.org.

## REFERENCES AND NOTES

1. Neumann M, Tolnay M, Mackenzie IRA, The molecular basis of frontotemporal dementia. Expert Rev. Mol. Med. 11, e23 (2009). [PubMed: 19638255]
2. Meneses A, Koga S, O'Leary J, Dickson DW, Bu G, Zhao N, TDP-43 pathology in Alzheimer's Disease. Neurodegeneration 16, 84 (2021).
3. Van Deerlin VM, Leverenz JB, Bekris LM, Bird TD, Yuan W, Elman LB, Clay D, Wood EMC, Chen-Plotkin AS, Martinez-Lage M, Steinbart E, McCluskey L, Grossman M, Neumann M, Wu IL, Yang WS, Kalb R, Galasko DR, Montine TJ, Trojanowski JQ, Lee VMY, Schellenberg GD, Yu CE, TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: A genetic and histopathological analysis. Lancet Neurol. 7, 409–416 (2008). [PubMed: 18396105]

4. Borroni B, Bonvicini C, Alberici A, Buratti E, Agosti C, Archetti S, Papetti A, Stuani C, Di Luca M, Gennarelli M, Padovani A, Mutation withinTARDBPleads to frontotemporal dementia without motor neuron disease. Hum. Mutat. 30, E974–E983 (2009). [PubMed: 19655382]
5. Halliday G, Bigio EH, Cairns NJ, Neumann M, MacKenzie IRA, Mann DMA, Mechanisms of disease in frontotemporal lobar degeneration: Gain of function versus loss of function effects. Acta Neuropathol. 124, 373–382 (2012). [PubMed: 22878865]
6. Vatsavayai SC, Yoon SJ, Gardner RC, Gendron TF, Vargas JNS, Trujillo A, Pribadi M, Phillips JJ, Gaus SE, Hixson JD, Garcia PA, Rabinovici GD, Coppola G, Geschwind DH, Petrucelli L, Miller BL, Seeley WW, Timing and significance of pathological features in C9orf72 expansion-associated frontotemporal dementia. Brain 139, 3202–3216 (2016). [PubMed: 27797809]
7. Yang C, Wang H, Qiao T, Yang B, Aliaga L, Qiu L, Tan W, Salameh J, McKenna-Yasek DM, Smith T, Peng L, Moore MJ, Brown RH, Cai H, Xu Z, Partial loss of TDP-43 function causes phenotypes of amyotrophic lateral sclerosis. Proc. Natl. Acad. Sci. U.S.A. 111, E1121–E1129 (2014). [PubMed: 24616503]
8. Iguchi Y, Katsuno M, Niwa JI, Takagi S, Ishigaki S, Ikenaka K, Kawai K, Watanabe H, Yamanaka K, Takahashi R, Misawa H, Sasaki S, Tanaka F, Sobue G, Loss of TDP-43 causes age-dependent progressive motor neuron degeneration. Brain 136, 1371–1382 (2013). [PubMed: 23449777]
9. Ling JP, Pletnikova O, Troncoso JC, Wong PC, TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science 349, 650–655 (2015). [PubMed: 26250685]
10. Humphrey J, Emmett W, Fratta P, Isaacs AM, Plagnol V, Quantitative analysis of cryptic splicing associated with TDP-43 depletion. BMC Med. Genomics 10, 38 (2017). [PubMed: 28549443]
11. Mehta PR, Brown A-L, Ward ME, Fratta P, The era of cryptic exons: Implications for ALS-FTD. Mol. Neurodegener. 18, 16 (2023). [PubMed: 36922834]
12. Prudencio M, Humphrey J, Pickles S, Brown A-L, Hill SE, Kachergus J, Shi J, Heckman M, Spiegel M, Cook C, Song Y, Yue M, Daughrity L, Carlomagno Y, Jansen-West K, Fernandez De Castro C, DeTure M, Koga S, Wang Y-C, Sivakumar P, Bodo C, Candalija A, Talbot K, Selvaraj BT, Burr K, Chandran S, Newcombe J, Lashley T, Hubbard I, Catalano D, Kim D, Propp N, Fennessey S, Fagegaltier D, Phatnani H, Secrier M, Fisher EMC, Oskarsson B, van Blitterswijk M, Rademakers R, Graff-Radford NR, Boeve B, Knopman DS, Petersen R, Josephs K, Thompson EA, Raj T, Ward ME, Dickson D, Gendron TF, Fratta P, Petrucelli L, Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. J. Clin. Invest. 130, 6080–6092 (2020). [PubMed: 32790644]
13. Klim JR, Williams LA, Limone F, Juan IGS, Davis-Dusenbery BN, Mordes DA, Burberry A, Steinbaugh MJ, Gamage KK, Kirchner R, Moccia R, Cassel SH, Chen K, Wainger BJ, Woolf CJ, Eggan K, ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. Nat. Neurosci. 22, 167–179 (2019). [PubMed: 30643292]
14. Melamed Z, López-Erauskin J, Baughn MW, Zhang O, Drenner K, Sun Y, Freyermuth F, McMahon MA, Beccari MS, Artates JW, Ohkubo T, Rodriguez M, Lin N, Wu D, Bennett CF, Rigo F, Da Cruz S, Ravits J, Lagier-Tourenne C, Cleveland DW, Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. Nat. Neurosci. 22, 180–190 (2019). [PubMed: 30643298]
15. Brown A-L, Wilkins OG, Keuss MJ, Hill SE, Zanovello M, Lee WC, Bampton A, Lee FCY, Masino L, Qi YA, Bryce-Smith S, Gatt A, Hallegger M, Fagegaltier D, Phatnani H, Phatnani H, Kwan J, Sareen D, Broach JR, Simmons Z, Arcila-Londono X, Lee EB, Van Deerlin VM, Shneider NA, Fraenkel E, Ostrow LW, Baas F, Zaitlen N, Berry JD, Malaspina A, Fratta P, Cox GA, Thompson LM, Finkbeiner S, Dardiotis E, Miller TM, Chandran S, Pal S, Hornstein E, MacGowan DJ, Heiman-Patterson T, Hammell MG, Patsopoulos NA, Butovsky O, Dubnau J, Nath A, Bowser R, Harms M, Aronica E, Poss M, Phillips-Cremins J, Crary J, Atassi N, Lange DJ, Adams DJ, Stefanis L, Gotkine M, Baloh RH, Babu S, Raj T, Paganoni S, Shalem O, Smith C, Zhang B, Harris B, Broce I, Drory V, Ravits J, McMillan C, Menon V, Wu L, Altschuler S, Lerner Y, Sattler R, Van Keuren-Jensen K, Rozenblatt-Rosen O, Lindblad-Toh K, Nicholson K, Gregersen P, Lee J-H, Kokos S, Muljo S, Newcombe J, Gustavsson EK, Seddighi S, Reyes JF, Coon SL, Ramos D, Schiavo G, Fisher EMC, Raj T, Secrier M, Lashley T, Ule J, Buratti E, Humphrey J, Ward ME, Fratta P, TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. Nature 603, 131–137 (2022). [PubMed: 35197628]

16. Ma XR, Prudencio M, Koike Y, Vatsavayai SC, Kim G, Harbinski F, Briner A, Rodriguez CM, Guo C, Akiyama T, Schmidt HB, Cummings BB, Wyatt DW, Kurylo K, Miller G, Mekhoubad S, Sallee N, Mekonnen G, Ganser L, Rubien JD, Jansen-West K, Cook CN, Pickles S, Oskarsson B, Graff-Radford NR, Boeve BF, Knopman DS, Petersen RC, Dickson DW, Shorter J, Myong S, Green EM, Seeley WW, Petrucelli L, Gitler AD, TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. Nature 603, 124–130 (2022). [PubMed: 35197626]
17. Prüss H, Autoantibodies in neurological disease. Rev. Immunol. 21, 798–813 (2021).
18. Macdonald IK, Parsy-Kowalska CB, Chapman CJ, Autoantibodies: Opportunities for early cancer detection. Trends Cancer 3, 198–213 (2017). [PubMed: 28718432]
19. Liu EY, Russ J, Cali CP, Phan JM, Amlie-Wolf A, Lee EB, Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. Cell Rep. 27, 1409–1421.e6 (2019). [PubMed: 31042469]
20. Hallegger M, Chakrabarti AM, Lee FCY, Lee BL, Amalietti AG, Odeh HM, Copley KE, Rubien JD, Portz B, Kuret K, Huppertz I, Rau F, Patani R, Fawzi NL, Shorter J, Luscombe NM, Ule J, TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. Cell 184, 4680–4696.e22 (2021). [PubMed: 34380047]
21. Baude A, Aaes TL, Zhai B, Al-Nakouzi N, Oo HZ, Daugaard M, Rohde M, Jäättelä M, Hepatoma-derived growth factor-related protein 2 promotes DNA repair by homologous recombination. Nucleic Acids Res. 44, 2214–2226 (2016). [PubMed: 26721387]
22. Barmada SJ, Skibinski G, Korb E, Rao EJ, Wu JY, Finkbeiner S, Cytoplasmic mislocalization of TDP-43 is toxic to neurons and enhanced by a mutation associated with familial amyotrophic lateral sclerosis. J. Neurosci. 30, 639 (2010). [PubMed: 20071528]
23. Yu H, Lu S, Gasior K, Singh D, Vazquez-Sanchez S, Tapia O, Toprani D, Beccari MS, Yates JR, Da Cruz S, Newby JM, Lafarga M, Gladfelter AS, Villa E, Cleveland DW, HSP70 chaperones RNA-free TDP-43 into anisotropic intranuclear liquid spherical shells. Science 371, eabb4309 (2021). [PubMed: 33335017]
24. Gasset-Rosa F, Lu S, Yu H, Chen C, Melamed Z, Guo L, Shorter J, Da Cruz S, Cleveland DW, Cytoplasmic TDP-43 de-mixing independent of stress granules drives inhibition of nuclear import, loss of nuclear tdp-43, and cell death. Neuron 102, 339–357. e7 (2019). [PubMed: 30853299]
25. Moisse K, Mepham J, Volkening K, Welch I, Hill T, Strong MJ, Cytosolic TDP-43 expression following axotomy is associated with caspase 3 activation in NFL−/− mice: Support for a role for TDP-43 in the physiological response to neuronal injury. Brain Res. 1296, 176–186 (2009). [PubMed: 19619516]
26. Moisse K, Volkening K, Leystra-Lantz C, Welch I, Hill T, Strong MJ, Divergent patterns of cytosolic TDP-43 and neuronal progranulin expression following axotomy: Implications for TDP-43 in the physiological response to neuronal injury. Brain Res. 1249, 202–211 (2009). [PubMed: 19046946]
27. Lee EB, Lee VMY, Trojanowski JQ, Gains or losses: Molecular mechanisms of TDP43-mediated neurodegeneration. Rev. Neurosci. 13, 38–50 (2011).
28. Graves MC, Fiala M, Dinglasan LAV, Liu NQ, Sayre J, Chiappelli F, van Kooten C, Vinters HV, Inflammation in amyotrophic lateral sclerosis spinal cord and brain is mediated by activated macrophages, mast cells and t cells. Amyotroph. Lateral Scler. Other Motor Neuron Disord. 5, 213–219 (2004). [PubMed: 15799549]
29. Engelhardt JI, Tajti J, Appel SH, Lymphocytic Infiltrates in the Spinal Cord in Amyotrophic Lateral Sclerosis. Arch. Neurol. 50, 30–36 (1993). [PubMed: 8093428]
30. Rodrigues Lima-Junior J, Sulzer D, Lindestam Arlehamn CS, Sette A, The role of immune-mediated alterations and disorders in ALS disease. Hum. Immunol. 82, 155–161 (2021). [PubMed: 33583639]
31. Gate D, Tapp E, Leventhal O, Shahid M, Nonninger TJ, Yang AC, Strempfl K, Unger MS, Fehlmann T, Oh H, Channappa D, Henderson VW, Keller A, Aigner L, Galasko DR, Davis MM, Poston KL, Wyss-Coray T, CD4+ T cells contribute to neurodegeneration in Lewy body dementia. Science 374, 868–874 (2021). [PubMed: 34648304]
32. Sulzer D, Alcalay RN, Garretti F, Cote L, Kanter E, Agin-Liebes J, Liong C, McMurtrey C, Hildebrand WH, Mao X, Dawson VL, Dawson TM, Oseroff C, Pham J, Sidney J, Dillon MB,

Carpenter C, Weiskopf D, Phillips E, Mallal S, Peters B, Frazier A, Lindestam Arlehamn CS, Sette A, T cells from patients with Parkinson's disease recognize α-synuclein peptides. Nature 546, 656–661 (2017). [PubMed: 28636593]

33. Bjornevik K, Cortese M, Healy BC, Kuhle J, Mina MJ, Leng Y, Elledge SJ, Niebuhr DW, Scher AI, Munger KL, Ascherio A, Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science 375, 296–301 (2022). [PubMed: 35025605]
34. Lanz TV, Brewer RC, Ho PP, Moon JS, Jude KM, Fernandez D, Fernandes RA, Gomez AM, Nadj GS, Bartley CM, Schubert RD, Hawes IA, Vazquez SE, Iyer M, Zuchero JB, Teegen B, Dunn JE, Lock CB, Kipp LB, Cotham VC, Ueberheide BM, Aftab BT, Anderson MS, DeRisi JL, Wilson MR, Bashford-Rogers RJM, Platten M, Garcia KC, Steinman L, Robinson WH, Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. Nature 603, 321–327 (2022). [PubMed: 35073561]
35. Josephs KA, Murray ME, Whitwell JL, Parisi JE, Petrucelli L, Jack CR, Petersen RC, Dickson DW, Staging TDP-43 pathology in Alzheimer's disease. Acta Neuropathol. 127, 441–450 (2014). [PubMed: 24240737]
36. Amador-Ortiz C, Lin WL, Ahmed Z, Personett D, Davies P, Duara R, Graff-Radford NR, Hutton ML, Dickson DW, TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer's disease. Ann. Neurol. 61, 435–445 (2007). [PubMed: 17469117]
37. Uryu K, Nakashima-Yasuda H, Forman MS, Kwong LK, Clark CM, Grossman M, Miller BL, Kretzschmar HA, Lee VMY, Trojanowski JQ, Neumann M, Concomitant TAR-DNA-binding protein 43 pathology is present in Alzheimer disease and corticobasal degeneration but not in other tauopathies. J. Neuropathol. Exp. Neurol. 67, 555–564 (2008). [PubMed: 18520774]
38. Picotti P, Aebersold R, Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. Methods 9, 555–566 (2012).
39. Rissin DM, Kan CW, Campbell TG, Howes SC, Fournier DR, Song L, Piech T, Patel PP, Chang L, Rivnak AJ, Ferrell EP, Randall JD, Provuncher GK, Walt DR, Duffy DC, Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. Nat. Biotechnol. 28, 595–599 (2010). [PubMed: 20495550]
40. Mao CP, Wang SC, Su YP, Tseng SH, He L, Wu AA, Roden RBS, Xiao J, Hung CF, Protein detection in blood with single-molecule imaging. Sci. Adv. 7, eabg6522 (2021). [PubMed: 34380620]
41. Tian R, Gachechiladze MA, Ludwig CH, Laurie MT, Hong JY, Nathaniel D, Prabhu AV, Fernandopulle MS, Patel R, Abshari M, Ward ME, Kampmann M, CRISPR interference-based platform for multimodal genetic screens in human iPSC-derived neurons. Neuron 104, 239–255.e12 (2019). [PubMed: 31422865]
42. Fernandopulle MS, Prestil R, Grunseich C, Wang C, Gan L, Ward ME, Transcription factor-mediated differentiation of human iPSCs into Neurons. Curr. Protoc. Cell Biol. 79, e51 (2018). [PubMed: 29924488]
43. Pantazis CB, Yang A, Lara E, McDonough JA, Blauwendraat C, Peng L, Oguro H, Kanaujiya J, Zou J, Sebesta D, Pratt G, Cross E, Blockwick J, Buxton P, Kinner-Bibeau L, Medura C, Tompkins C, Hughes S, Santiana M, Faghri F, Nalls MA, Vitale D, Ballard S, Qi YA, Ramos DM, Anderson KM, Stadler J, Narayan P, Papademetriou J, Reilly L, Nelson MP, Aggarwal S, Rosen LU, Kirwan P, Pisupati V, Coon SL, Scholz SW, Priebe T, Öttl M, Dong J, Meijer M, Janssen LJM, Lourenco VS, van der Kant R, Crusius D, Paquet D, Raulin A-C, Bu G, Held A, Wainger BJ, Gabriele RMC, Casey JM, Wray S, Abu-Bonsrah D, Parish CL, Beccari MS, Cleveland DW, Li E, Rose IVL, Kampmann M, Calatayud Aristoy C, Verstreken P, Heinrich L, Chen MY, Schüle B, Dou D, Holzbaur ELF, Zanellati MC, Basundra R, Deshmukh M, Cohen S, Khanna R, Raman M, Nevin ZS, Matia M, Van Lent J, Timmerman V, Conklin BR, Johnson Chase K, Zhang K, Funes S, Bosco DA, Erlebach L, Welzer M, Kronenberg-Versteeg D, Lyu G, Arenas E, Coccia E, Sarrafha L, Ahfeldt T, Marioni JC, Skarnes WC, Cookson MR, Ward ME, Merkle FT, A reference human induced pluripotent stem cell line for large-scale collaborative studies. Cell Stem Cell 29, 1685–1702.e22 (2022). [PubMed: 36459969]
44. Martin M, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17, (2011).

45. Andrews S, FastQC - A quality control tool for high throughput sequence data; www.bioinformatics.babraham.ac.uk/projects/fastqc/. Babraham Bioinforma (2010).
46. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR, STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]
47. Love MI, Huber W, Anders S, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). [PubMed: 25516281]
48. Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y, A new view of transcriptome complexity and regulation through the lens of local splicing variations. eLife 5, e11752 (2016). [PubMed: 26829591]
49. Li H, Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018). [PubMed: 29750242]
50. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J, Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat. Methods 20, 1187–1195 (2023). [PubMed: 37308696]
51. Wilkins OG, Ule J, Ribocutter: Cas9-mediated rRNA depletion from multiplexed Ribo-seq libraries. bioRxiv 2021.07.14.451473v1.full (2021). www.biorxiv.org/content/10.1101/2021.07.14.451473v1.full.
52. Zhang D, Reynolds RH, Garcia-Ruiz S, Gustavsson EK, Sethi S, Aguti S, Barbosa IA, Collier JJ, Houlden H, McFarland R, Muntoni F, Oláhová M, Poulton J, Simpson M, Pitceathly RDS, Taylor RW, Zhou H, Deshpande C, Botia JA, Collado-Torres L, Ryten M, Detection of pathogenic splicing events from RNA-sequencing data using dasper. bioRxiv 2021.03.29.437534v1 (2021). www.biorxiv.org/content/10.1101/2021.03.29.437534v1.
53. Shao M, Kingsford C, Accurate assembly of transcripts through phase-preserving graph decomposition. Nat. Biotechnol. 35, 1167–1169 (2017). [PubMed: 29131147]
54. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M, Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20, 278 (2019). [PubMed: 31842956]
55. Reilly L, Peng L, Lara E, Ramos D, Fernandopulle M, Pantazis CB, Stadler J, Santiana M, Dadu A, Iben J, Faghri F, Nalls MA, Coon SL, Narayan P, Singleton AB, Cookson MR, Ward ME, Qi YA, A fully automated FAIMS-DIA proteomic pipeline for high-throughput characterization of iPSC-derived neurons. bioRxiv 2021.11.24.469921v1 (2021). www.biorxiv.org/content/10.1101/2021.11.24.469921v1.
56. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B, PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. Mol. Cell. Proteomics 11, M111.010587 (2012).
57. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ, Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966–968 (2010). [PubMed: 20147306]
58. Martinez-Val A, Bekker-Jensen DB, Hogrebe A, Olsen JV, Data processing and analysis for DIA-based phosphoproteomics using spectronaut. Methods Mol. Biol. 2361, 95–107 (2021). [PubMed: 34236657]
59. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M, pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77 (2011). [PubMed: 21414208]
60. Ge SX, Jung D, Jung D, Yao R, ShinyGO: A graphical gene-set enrichment tool for animals and plants. Bioinformatics 36, 2628–2629 (2020). [PubMed: 31882993]
61. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Von Mering C, STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613 (2019). [PubMed: 30476243]
62. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T, Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003). [PubMed: 14597658]

63. Higginbotham L, Ping L, Dammer EB, Duong DM, Zhou M, Gearing M, Hurst C, Glass JD, Factor SA, Johnson ECB, Hajjar I, Lah JJ, Levey AI, Seyfried NT, Integrated proteomics reveals brain-based cerebrospinal fluid biomarkers in asymptomatic and symptomatic Alzheimer's disease. Sci. Adv. 6, eaaz9360 (2020). [PubMed: 33087358]
64. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J, Sustainable data analysis with Snakemake. F1000Res 10, 33 (2021). [PubMed: 34035898]
65. Quinlan AR, Hall IM, BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). [PubMed: 20110278]
66. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO, The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596 (2013). [PubMed: 23193283]
67. Langmead B, Salzberg SL, Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). [PubMed: 22388286]
68. Bairoch A, Apweiler R, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45–48 (2000). [PubMed: 10592178]
69. Johnson ECB, Dammer EB, Duong DM, Ping L, Zhou M, Yin L, Higginbotham LA, Guajardo A, White B, Troncoso JC, Thambisetty M, Montine TJ, Lee EB, Trojanowski JQ, Beach TG, Reiman EM, Haroutunian V, Wang M, Schadt E, Zhang B, Dickson DW, Ertekin-Taner N, Golde TE, Petyuk VA, De Jager PL, Bennett DA, Wingo TS, Rangaraju S, Hajjar I, Shulman JM, Lah JJ, Levey AI, Seyfried NT, Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nat. Med. 26, 769–780 (2020). [PubMed: 32284590]
70. Tam OH, Rozhkov NV, Shaw R, Kim D, Hubbard I, Fennessey S, Propp N, Phatnani H, Kwan J, Sareen D, Broach JR, Simmons Z, Arcila-Londono X, Lee EB, Van Deerlin VM, Shneider NA, Fraenkel E, Ostrow LW, Baas F, Zaitlen N, Berry JD, Malaspina A, Fratta P, Cox GA, Thompson LM, Finkbeiner S, Dardiotis E, Miller TM, Chandran S, Pal S, Hornstein E, MacGowan DJ, Heiman-Patterson T, Hammell MG, Patsopoulos NA, Butovsky O, Dubnau J, Nath A, Bowser R, Harms M, Aronica E, Poss M, Phillips-Cremins J, Crary J, Atassi N, Lange DJ, Adams DJ, Stefanis L, Gotkine M, Baloh R, Babu S, Raj T, Paganoni S, Shalem O, Smith C, Zhang B, Harris BT, Fagegaltier D, Ravits J, Dubnau J, Hammell MG, Postmortem cortex samples identify distinct molecular subtypes of ALS: Retrotransposon activation, oxidative stress, and activated Glia. Cell Rep. 29, 1164–1177.e5 (2019). [PubMed: 31665631]
71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]
72. Prudencio M, Gonzales PK, Cook CN, Gendron TF, Daughrity LM, Song Y, Ebbert MTW, van Blitterswijk M, Zhang YJ, Jansen-West K, Baker MC, DeTure M, Rademakers R, Boylan KB, Dickson DW, Petrucelli L, Link CD, Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. Hum. Mol. Genet. 26, 3421–3431 (2017). [PubMed: 28637276]
73. Zhang Y-J, Gendron TF, Grima JC, Sasaguri H, Jansen-West K, Xu Y-F, Katzman RB, Gass J, Murray ME, Shinohara M, Lin W-L, Garrett A, Stankowski JN, Daughrity L, Tong J, Perkerson EA, Yue M, Chew J, Castanedes-Casey M, Kurti A, Wang ZS, Liesinger AM, Baker JD, Jiang J, Lagier-Tourenne C, Edbauer D, Cleveland DW, Rademakers R, Boylan KB, Bu G, Link CD, Dickey CA, Rothstein JD, Dickson DW, Fryer JD, Petrucelli L, C9ORF72 poly(GA) aggregates sequester and impair HR23 and nucleocytoplasmic transport proteins. Nat. Neurosci. 19, 668–677 (2016). [PubMed: 26998601]

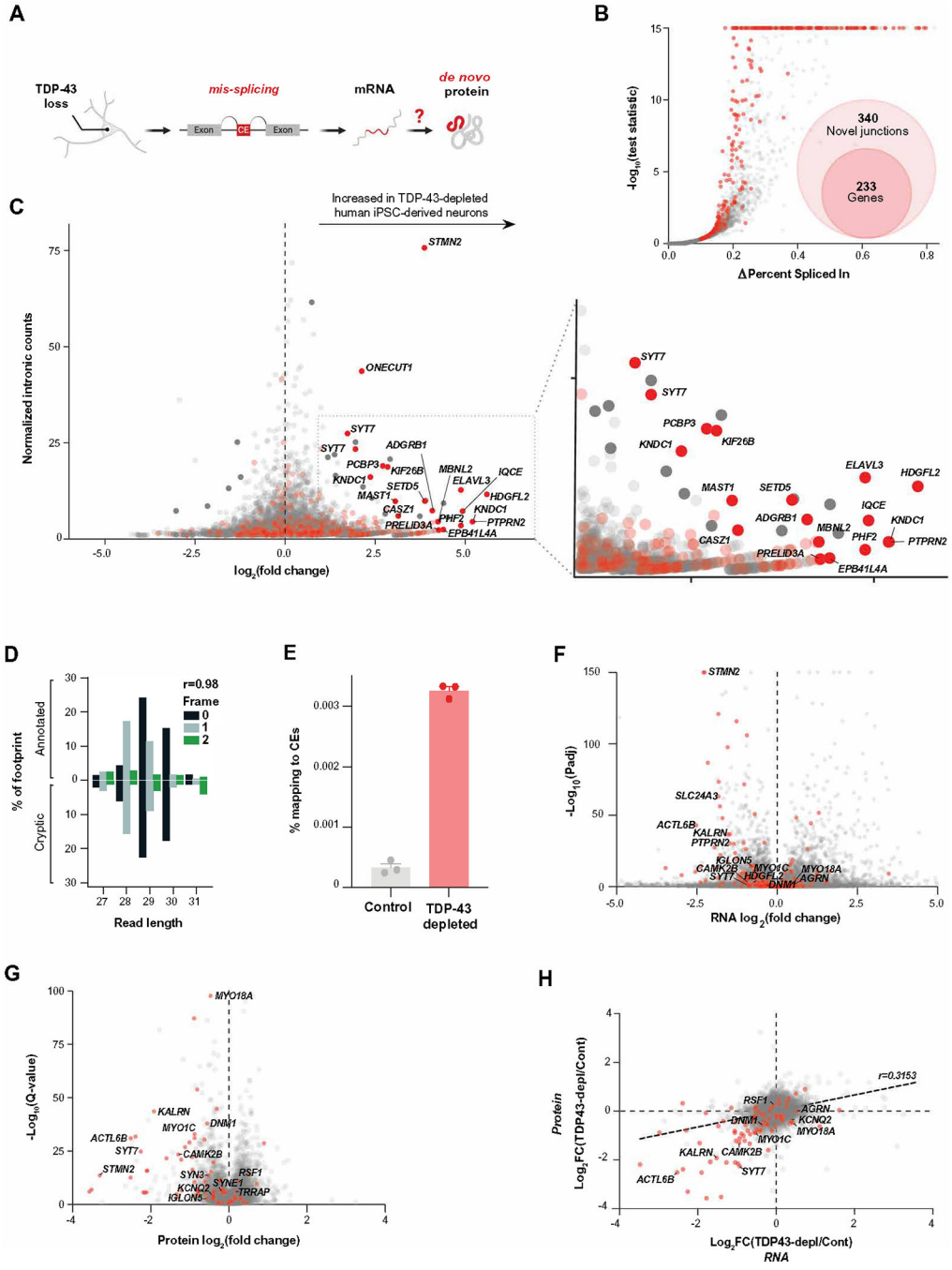


**Fig. 1. Transcriptional and proteomic analysis of TDP-43–depleted human iPSC–derived neurons.**

(**A**) Overview of CE RNA and potential de novo protein production due to pathological mis-splicing in TDP-43–depleted human iPSC–derived neurons. (**B**) Shown is differential splicing in TDP-43–depleted human iPSC–derived neurons compared with human iPSC–derived control neurons expressing normal amounts of TDP-43. CE genes are shown in red. (**C**) Differential ribosome footprint density of intronic mRNA is shown for TDP-43–depleted human iPSC–derived neurons compared with human iPSC–derived

control neurons. CE genes are shown in red. (**D**) Shown is the percentage abundance of each footprint type, defined by length and sub-codon position, for annotated coding sequences (positive $Y$ axis) and a subset of high-confidence CEs, which are predicted to be translated (negative $Y$ axis). Pearson's correlation between the values for annotated and cryptic transcripts is shown. Data are from one TDP-43–depleted human iPSC–derived neuronal cell line. (**E**) Shown is the percentage of footprints aligning to the subset of CEs used in (C) for three TDP-43–depleted and three control human iPSC–derived neuronal samples. (**F**) Differential transcript abundance was quantified in TDP-43–depleted and control human iPSC–derived neurons using total short-read RNA-seq. CE genes are shown in red. (**G**) Differential protein abundance in TDP-43–depleted and control human iPSC–derived neurons was quantified using mass spectrometry proteomics. CE genes are shown in red. (**H**) Quantification of mRNA (RNA-seq) and protein by data-independent acquisition (DIA) proteomics of TDP-43–depleted and control human iPSC–derived neurons is shown. CEs are shown in red.

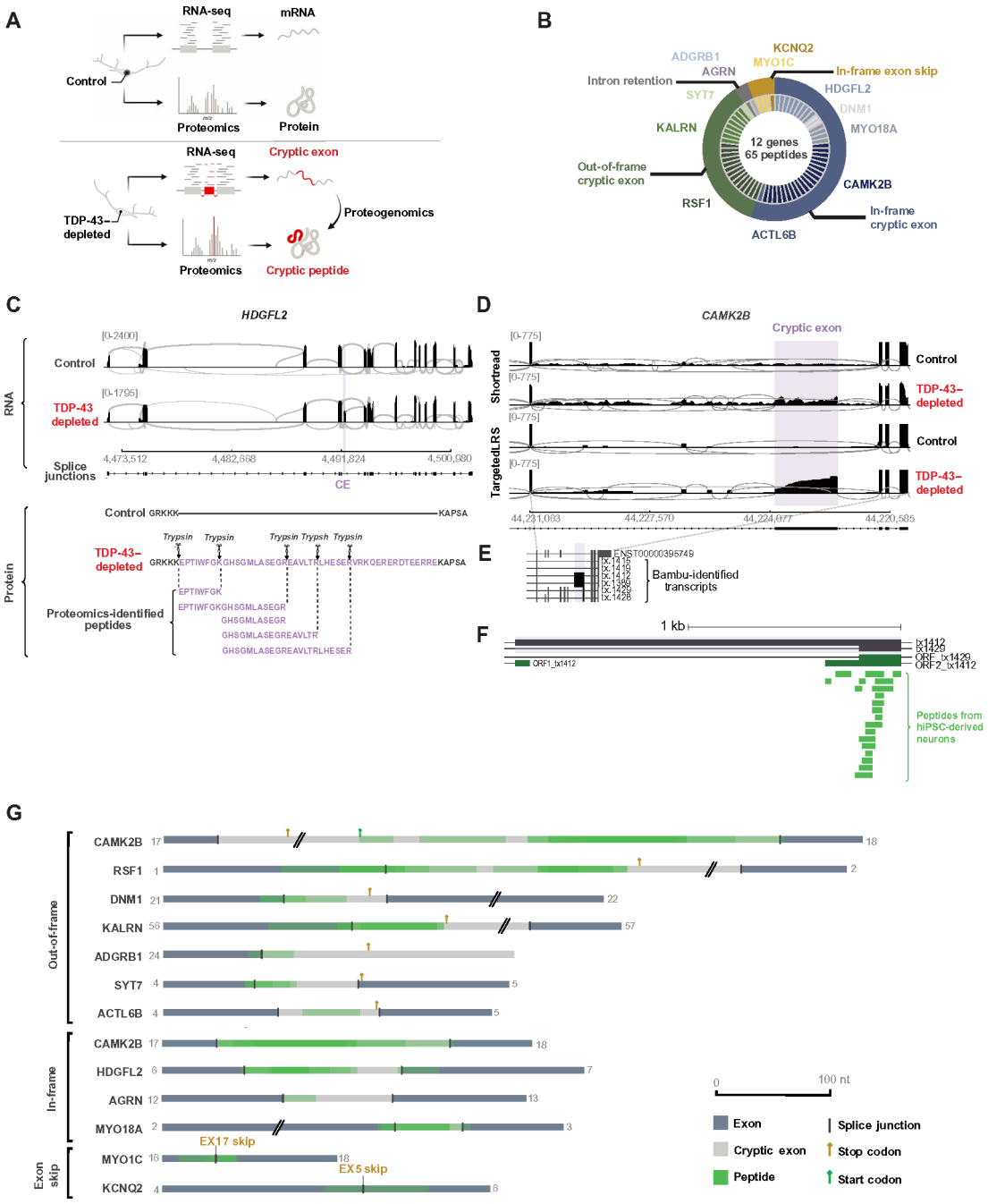


**Fig. 2. Formation of de novo cryptic peptides from mis-spliced transcripts in TDP-43–depleted human iPSC–derived neurons.**

(**A**) Proteogenomic pipeline used to identify de novo cryptic peptides caused by RNA mis-splicing in TDP-43–depleted human iPSC–derived neurons. (**B**) Proteogenomic analysis of TDP-43–depleted human iPSC–derived neurons identified 65 putative trypsin-digested cryptic peptides across 12 genes. The outer circle of the pie chart represents genes. The inner circle represents the number of putative trypsin-digested cryptic peptides that were identified for each gene. (**C**) Representative sashimi plots showing the inclusion of an in-frame CE (purple) in HDGFL2 transcripts in TDP-43–depleted human iPSC–derived

neurons. The amino acid sequence of the putative translation of this CE is shown below the sashimi plot, with trypsin cleavage sites and proteogenomic-identified cryptic peptides annotated. (**D**) Representative sashimi plots comparing Illumina short-read RNA-seq versus Nanopore long-read sequencing of the *CAMK2B* transcript in control and TDP-43–depleted human iPSC–derived neurons. The CE expressed in TDP-43–depleted human iPSC–derived neurons is highlighted in purple. (**E**) Precise mapping of *CAMK2B* exon junctions and splice isoforms of transcripts in TDP-43–depleted human iPSC–derived neurons using Nanopore long-read sequencing. (**F**) Tryptic-cryptic peptides (green) identified using the proteogenomic pipeline are mapped to an in-frame CE in *CAMK2B* identified using Nanopore long-read sequencing. (**G**) Graphical representation of proteogenomic-identified cryptic peptides mapped to transcripts from Nanopore long-read sequencing. Two genes, *MYO1C* and *KCNQ2*, contained an in-frame exon skipping event upon TDP-43 loss. Four genes, *HDGFL2*, *AGRN*, *MYO18A*, and *CAMK2B*, contained in-frame CEs. An out-of-frame *CAMK2B* CE was also identified. Canonical upstream and downstream exons are in dark gray. CEs are in light gray. Cryptic peptide locations are overlaid in green.

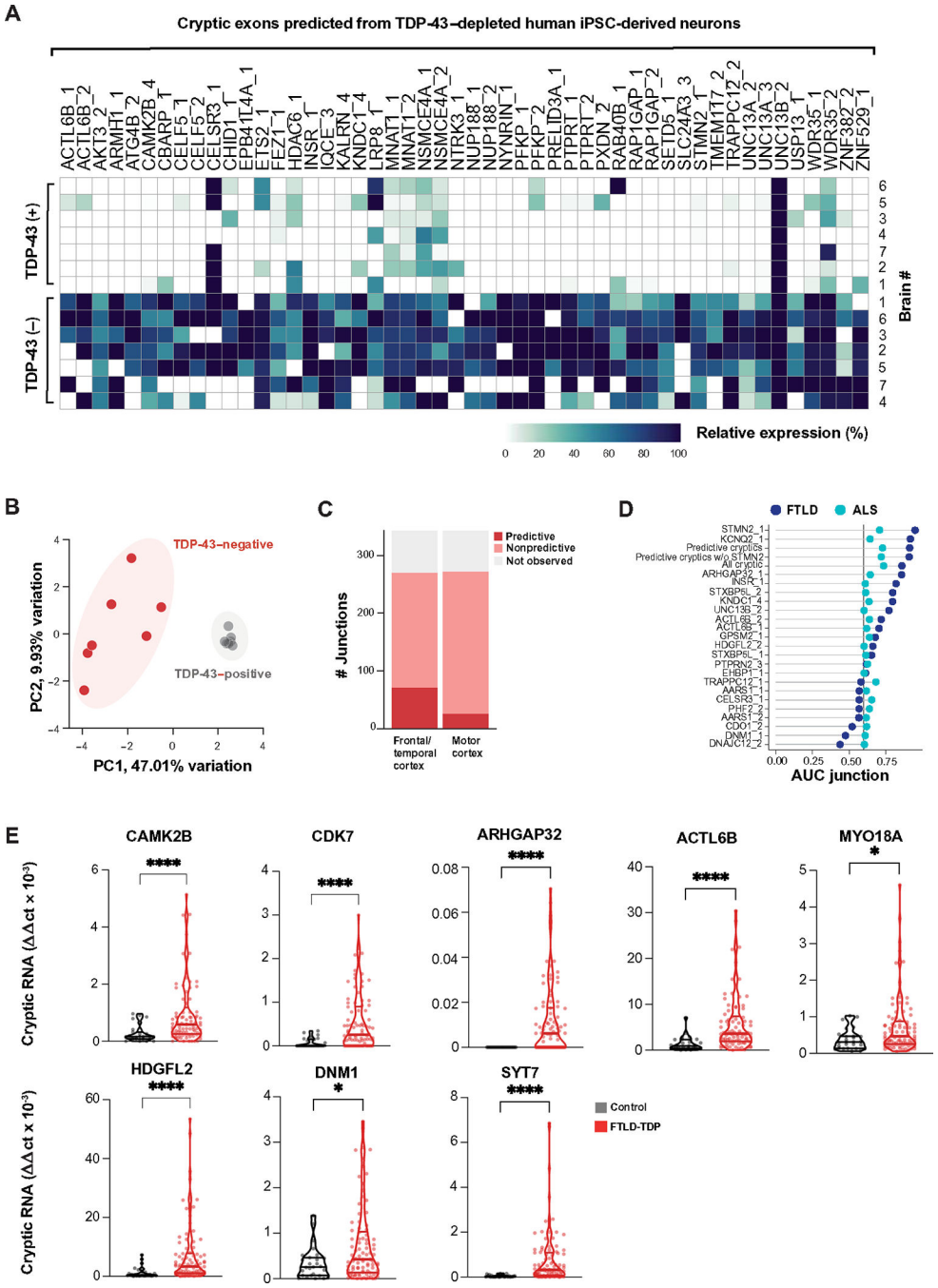


**Fig. 3. TDP-43 cryptic exons in TDP-43–depleted human iPSC–derived neurons predict TDP-43 pathology in postmortem brain tissue.**

(**A**) Heatmap showing the relative abundance of CEs predicted from TDP-43–depleted human iPSC–derived neurons in FACS-sorted TDP-43–positive and TDP-43–negative neuronal nuclei preparations from postmortem ALS/FTD cortical samples ($n = 7$) (17). The "percent spliced in" (PSI) value represents the ratio of transcripts that include a splicing event versus the total number of transcripts. We defined enriched junctions as those with PSI in TDP-43–negative nuclei greater than twice the PSI in TDP-43–positive nuclei (mean TDP-43–negative PSI > 0.10). The top 50 most expressed cryptic splice junctions

in TDP-43–negative nuclei, compared with TDP-43–positive nuclei, are shown, and cases were organized by unsupervised hierarchical clustering based on CE PSI. (**B**) Principal components analysis (PCA) on PSI of 230 predicted TDP-43 CEs in TDP-43–positive and TDP-43–negative neuronal nuclei from ALS/FTD postmortem cortical samples. (**C**) Bar plot showing the number of predicted CEs detected in bulk RNA-seq of ALS/FTD postmortem brain tissue samples from the NYGC biobank. CEs were classified by whether they were observed in postmortem brain tissue and were detected/nonpredictive or detected/predictive of TDP-43 pathology versus non–TDP-43 pathology. (**D**) AUC is shown for predicted CEs that identified TDP-43 pathology in FTLD frontal/temporal cortex and ALS motor cortex postmortem tissue from patients with ALS/FTLD patients. Meta-expression scores for all CEs, only predictive exons (AUC > = 0.6), and predictive exons excluding *STMN2* expression are shown. (**E**) Shown is quantitative RT-PCR–based validation of eight CEs in an independent set of postmortem frontal cortex brain samples from patients with FTLD-TDP ($n = 89$) and age-matched controls with non-neurological disease ($n = 27$). Data are presented as means ± SEM. *P* values are from the Mann-Whitney test: $*P < 0.05$ and $****P \leq 0.0001$.

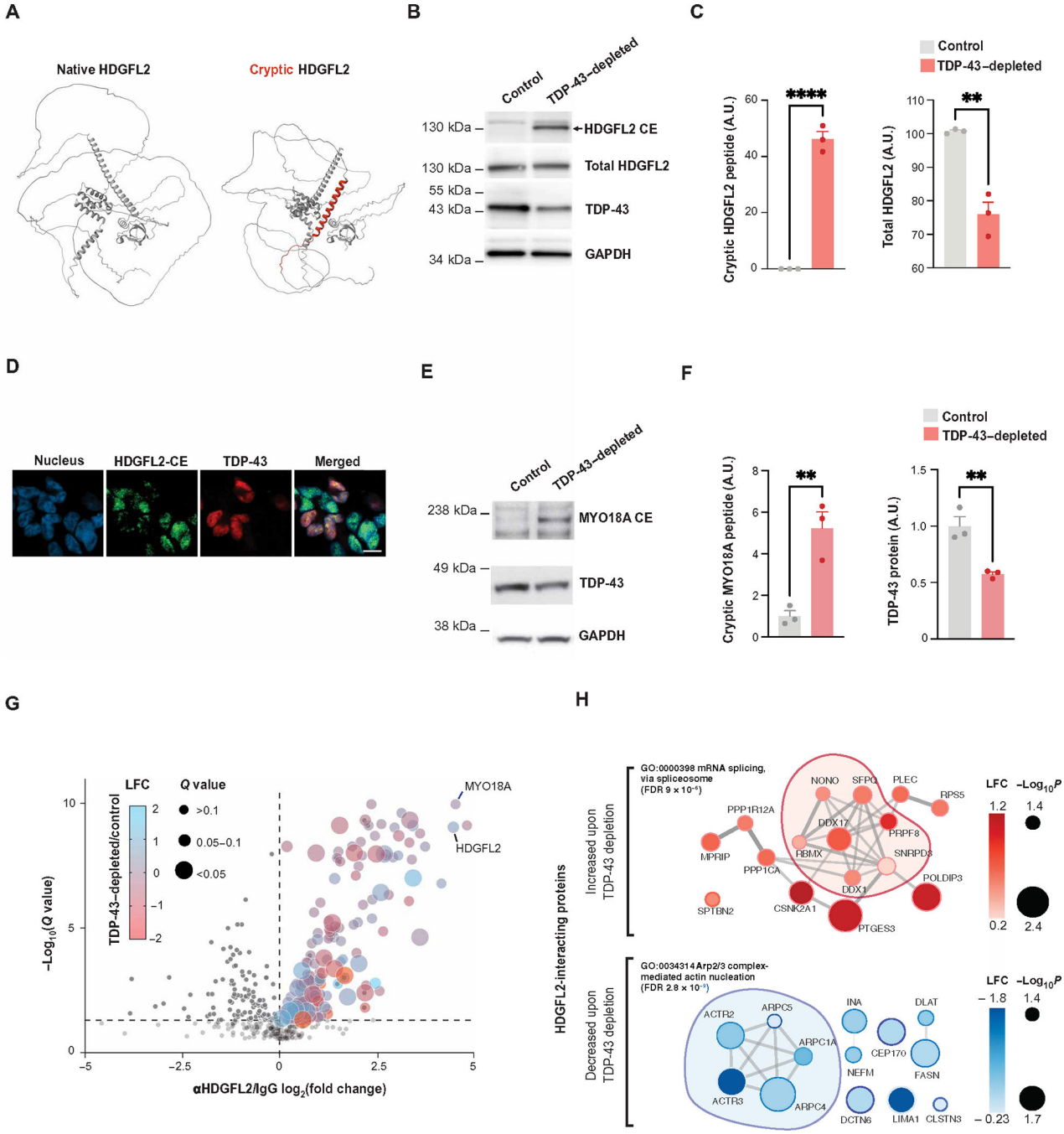


**Fig. 4. Cryptic peptides in TDP-43–depleted human iPSC–derived neurons alter the protein interactome.**

(**A**) Predicted structure of HDGFL2 (using AlphaFold), annotated with the predicted cryptic peptide induced by TDP-43 depletion highlighted in red. (**B**) Antibody-based detection of an HDGFL2 cryptic peptide in TDP-43–depleted human iPSC–derived neurons. Representative Western blot shows a band of the expected molecular weight of HDGFL2-CE specifically in TDP-43–depleted but not control human iPSC–derived neurons. (**C**) Quantification of HDGFL2-CE and total HDGFL2 in Western blot from (B) ($n = 3$, two-sample $t$ test, $^{**}P$

< 0.01 and ****$P$ < 0.0001; Shapiro-Wilk test for normality, $P$ > 0.05 not significant). (**D**) Immunofluorescence staining highlights the selective expression of the HDGFL2 cryptic peptide in TDP-43–depleted human iPSC–derived neurons (scale bar,10 μm). (**E**) Antibody-based detection of a MYO18A cryptic peptide in TDP-43–depleted human iPSC–derived neurons. Representative Western blot shows a band of the expected molecular weight of the MYO18A-CE specifically in TDP-43–depleted human iPSC–derived neurons. (**F**) Quantification of MYO18A-CE and TDP-43 protein expression ($n$ = 3, two-sample $t$ test, **$P$ < 0.01; Shapiro-Wilk test for normality, $P$ > 0.05 not significant). (**G**) Affinity purification mass spectrometry analysis of HDGFL2 protein-protein interactions in TDP-43–depleted and control human iPSC–derived neurons. A volcano plot of coimmunoprecipitated proteins using anti-HDGFL2 antibody versus control IgG is shown. The dot color reflects log fold change (LFC) in TDP-43–depleted versus control human iPSC–derived neurons, and the dot size reflects the adjusted $P$ value ($q$ value). (**H**) STRING diagram of proteins whose interactions with HDGFL2 were significantly altered by TDP-43–depleted versus control human iPSC–derived neurons ($P_{adj}$ < 0.05). The dot color reflects LFC in TDP-43–depleted versus control human iPSC–derived neurons.

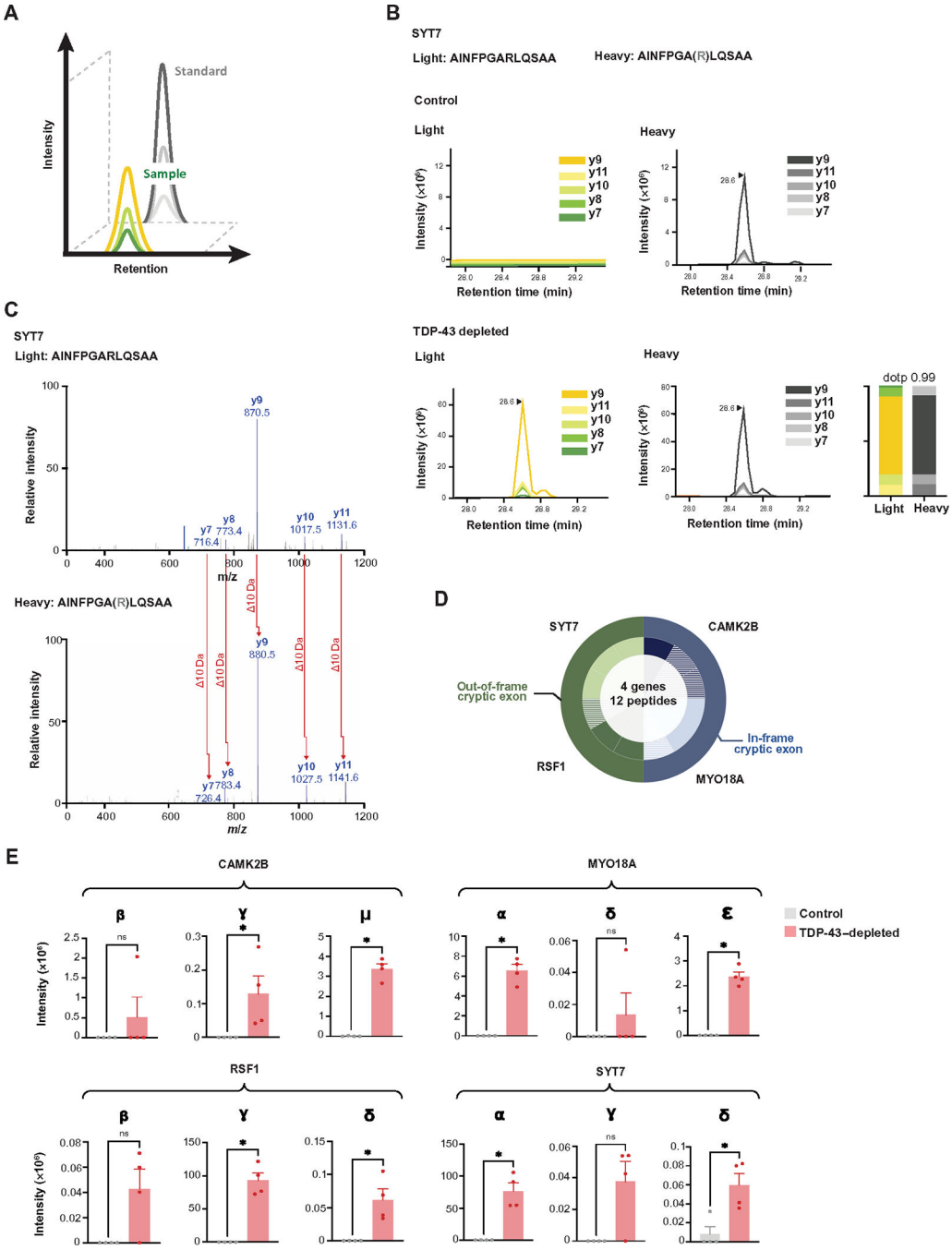


**Fig. 5. Scalable cryptic peptide validation in TDP-43–depleted human iPSC–derived neurons by targeted proteomics.**

(**A**) Schematic of PRM-MS. Co-elution of SIL peptides allows for sensitive measurement of corresponding endogenous peptides. (**B**) PRM-MS assay using a synthetic SIL peptide internal standard identifies a cryptic peptide in SYT7 in TDP-43–depleted but not in control human iPSC–derived neurons. The spectral plot of heavy standards and light (endogenous) y ions from an SYT7 cryptic peptide is shown, with accompanying dot product (dotp), which indicates the correlation between the peptide fragment-ion peak areas and theoretical

spectra. (**C**) Corresponding mass spectra of endogenous and heavy peptide standards of the SYT7 cryptic peptide in TDP-43–depleted human iPSC–derived neurons. (**D**) The detection of 12 trypsin-digested cryptic peptides across four genes using single-shot PRM assays in TDP-43–depleted human iPSC–derived neuronal lysates is shown. The outer circle represents the gene, and the inner circle represents the number of cryptic peptides detected by PRM per gene. Hatched color signifies the successful detection of one to two y ions; solid color signifies the detection of three or more y ions. (**E**) Quantification of cryptic peptide expression in TDP-43–depleted and control human iPSC–derived neurons using PRM assays. $n = 4$ replicates per sample. Mann-Whitney $U$ test. $*P < 0.05$, ns, not significant.

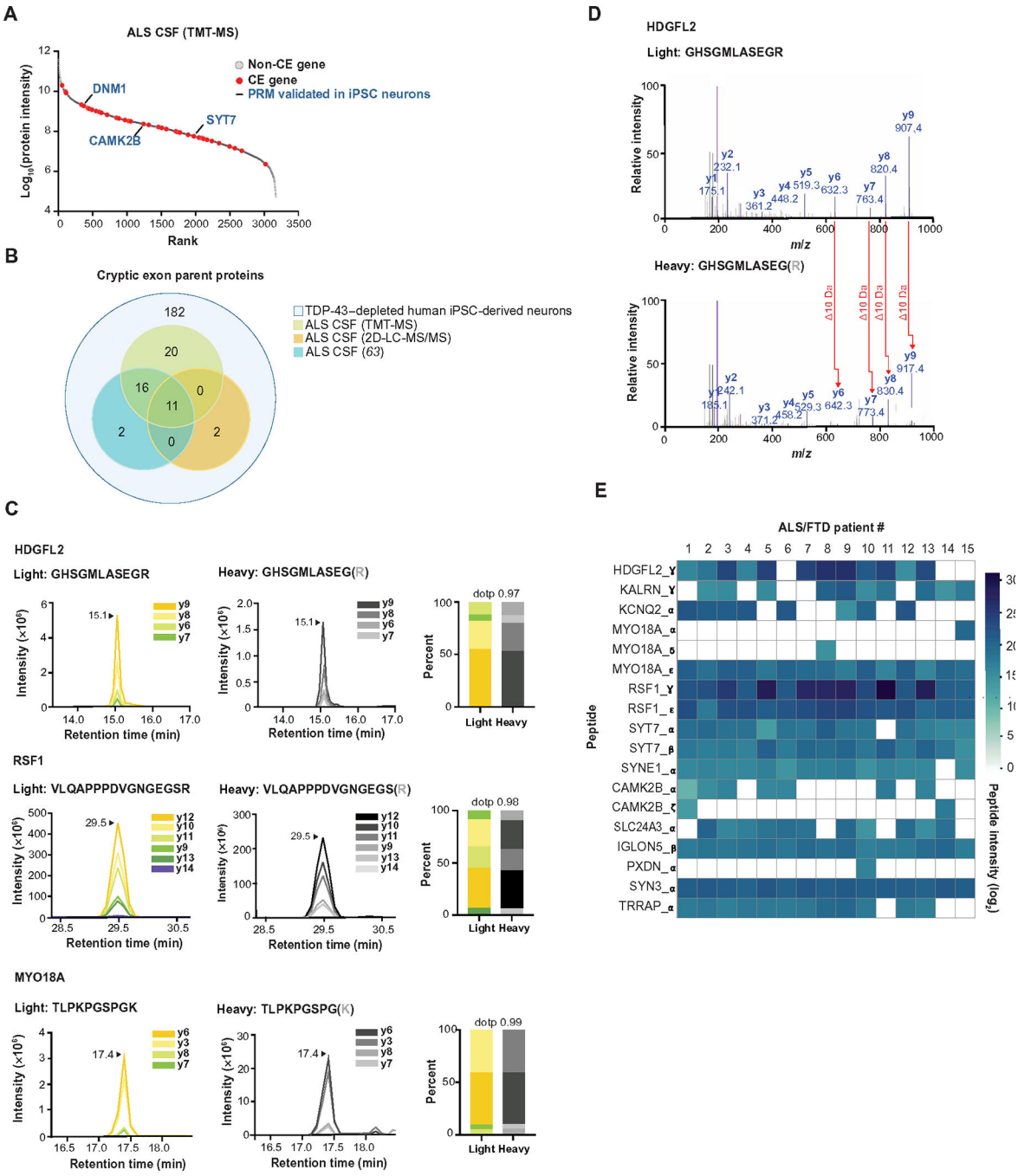


**Fig. 6. Cryptic peptides are present in CSF samples from patients with ALS.**

(**A**) Rank plot of proteins detected in CSF samples from 24 patients with ALS. Forty-seven CE genes that were predicted in TDP-43–depleted human iPSC–derived neurons are shown in red. PRM-validated cryptic peptides expressing genes (from human iPSC–derived neuron studies) are annotated in blue text. (**B**) Euler diagram of three CSF proteomics datasets versus CEs predicted in TDP-43–depleted human iPSC–derived neurons. (**C**) Representative spectra of three heavy (standard) and light (endogenous) cryptic peptides detected in CSF from patients with ALS/FTD spectrum disorders (*n* = 13 ALS, *n* = 1 ALS/FTD, *n* = 1 ALS/

mild cognitive impairment). Also shown is the dotp, which indicates the correlation between the peptide fragment-ion peak areas and theoretical spectra. The spectra for 15 additional cryptic peptides in ALS/FTD CSF samples are shown in fig. S9A. (**D**) MS/MS spectrum of a cryptic peptide in HDGFL2 that corresponds to the reference peptide (top) and endogenous peptide (bottom) detected in ALS/FTD CSF samples. (**E**) Heatmap of AUC intensities of 18 cryptic peptides from 13 different proteins in CSF samples from 15 patients with ALS/FTD.