TDP-43 loss induces extensive cryptic polyadenylation in ALS/FTD

Sam Bryce-Smith[1], Anna-Leigh Brown[1], Puja R. Mehta[1], Francesca Mattedi[1], Alla Mikheenko[1], Simone Barattucci[1], Matteo Zanovello[1], Dario Dattilo[1], Matthew Yome[1], Sarah E. Hill[2], Yue A. Qi[2], Oscar G. Wilkins[1,3], Kai Sun[1], Eugeni Ryadnov[1], Yixuan Wan[1], NYGC ALS Consortium, Jose Norberto S. Vargas[1], Nicol Birsa[1], Towfique Raj[4,5,6,7], Jack Humphrey[4,5,6,7], Matthew Keuss[1], Michael Ward[2], Maria Secrier[8,*] and Pietro Fratta[1,3*]

Affiliations:

1. *UCL Queen Square Motor Neuron Disease Centre, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, UCL, London, UK*
2. *National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD, USA*
3. *The Francis Crick Institute, London, UK*
4. *Nash Family Department of Neuroscience & Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.*
5. *Ronald M. Loeb Center for Alzheimer's Disease, Icahn School of Medicine at Mount Sinai, New York, NY, USA.*
6. *Department of Genetics and Genomic Sciences & Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.*
7. *Estelle and Daniel Maggin Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.*
8. *UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, London, UK*

** correspondence to p.fratta@ucl.ac.uk and m.secrier@ucl.ac.uk*

Abstract

Nuclear depletion and cytoplasmic aggregation of the RNA-binding protein TDP-43 is the hallmark of ALS, occurring in over 97% of cases. A key consequence of TDP-43 nuclear loss is the de-repression of cryptic exons. Whilst TDP-43 regulated cryptic splicing is increasingly well catalogued, cryptic alternative polyadenylation (APA) events, which define the 3' end of last exons, have been largely overlooked, especially when not associated with novel upstream splice junctions. We developed a novel bioinformatic approach to reliably identify distinct APA event types: alternative last exons (ALE), 3'UTR extensions (3'Ext) and intronic polyadenylation (IPA) events. We identified novel neuronal cryptic APA sites induced by TDP-43 loss of function by systematically applying our pipeline to a compendium of publicly available and in house datasets. We find that TDP-43 binding sites and target motifs are enriched at these cryptic

events and that TDP-43 can have both repressive and enhancing action on APA. Importantly, all categories of cryptic APA can also be identified in ALS and FTD post mortem brain regions with TDP-43 proteinopathy underlining their potential disease relevance. RNA-seq and Ribo-seq analyses indicate that distinct cryptic APA categories have different downstream effects on transcript and translation. Intriguingly, cryptic 3'Exts occur in multiple transcription factors, such as *ELK1*, *SIX3,* and *TLX1,* and lead to an increase in wild-type protein levels and function. Finally, we show that an increase in RNA stability leading to a higher cytoplasmic localisation underlies these observations. In summary, we demonstrate that TDP-43 nuclear depletion induces a novel category of cryptic RNA processing events and we expand the palette of TDP-43 loss consequences by showing this can also lead to an increase in normal protein translation.

Introduction

Cytoplasmic aggregates and nuclear depletion of TDP-43 are pathological hallmarks of a spectrum of neurodegenerative diseases, including over 97% of amyotrophic lateral sclerosis (ALS) cases[1], 45% of frontotemporal dementia (FTD)[2] and over 50% of Alzheimer's disease cases[3]. Under normal conditions, TDP-43 is a predominantly nuclear protein with multiple roles in regulation of RNA processing and metabolism, including alternative splicing, alternative polyadenylation (APA)[4–6], and transport[7]. Significant attention has been drawn to TDP-43's ability to repress the inclusion of pre-mRNA sequences in mature transcripts[8]: loss of nuclear TDP-43 leads to the inclusion of 'cryptic' exons in mature transcripts both *in vitro* and in post mortem tissue[7], contributing to disease progression[10,11]. Cryptic exons can lead to protein loss through RNA degradation by nonsense mediated decay[12], or can be translated to produce cryptic peptides[13,14].

Cleavage and polyadenylation defines the 3′end of last exons and subsequently mature transcripts[15]. Up to 70% of human protein-coding and long non-coding RNA genes can undergo polyadenylation at multiple locations in the gene body (alternative polyadenylation, APA), and can be subdivided in three main category of events: alternative last exons (ALE), 3'UTR extensions (3'Ext) and intronic polyadenylation events (IPA). In alternative last exons (ALE), the polyA usage is determined by an upstream alternative splice junction which defines an alternative last exon. In 3'Ext events, APA sites are independent of splice junctions and occur within 3'UTR regions and affect 3'UTR sequence and length, which is implicated in the regulation of transcript stability, localisation and translation[16]. Finally in IPA events, APA occurs within introns giving rise to transcripts with different protein coding potential and can affect full-length protein dosage[17,18].

TDP-43 regulated cryptic APA has not been systematically explored in a neuronal context. Here, we report widespread cryptic APA upon TDP-43 depletion in cell models, including events which were not previously detected with conventional splicing analyses. A substantial number are expressed in post-mortem ALS & ALS/FTD tissue with TDP-43 loss, underlining their potential involvement in pathogenic mechanisms and/or utility as biomarkers of TDP-43 pathology. We focus on a novel class of 3'Ext APA and show they can lead to increased translation levels. Moreover, we use metabolic labelling to demonstrate that such cryptic 3'Ext are associated with increased RNA stability, and in the case of *ELK1*, coincide with increased cytoplasmic RNA localisation. Our data therefore identifies a novel consequence for cryptic RNA processing, and shows that in addition to leading to protein reduction or the formation of altered proteins, this can also lead to overexpression of normal proteins, and an increase in their function.

Results

Identification of cryptic alternative polyadenylation events induced by TDP-43 loss

While TDP-43's role in regulating APA and cryptic splicing is well-known, cryptic APA occurring upon TPD-43 loss-of-function has yet to be explored. In order to comprehensively address this question, we curated a compendium of publicly available and newly generated bulk RNA-seq datasets with TDP-43 depletion (**Supplementary Table 1**). We assembled a computational pipeline to identify novel last exons from RNA-seq data, which defines last exon frames using StringTie[19], and then filters and categorises as spurious predicted 3'ends lacking the presence of reference polyA sites[20] or a conserved polyA signal hexamer[21] (**Fig. 1A**). Isoform level quantification was performed using Salmon[22], and differential usage between experimental conditions was assessed using DEXSeq[23].

We subdivided our events into three main categories: ALEs, IPAs and 3'Ext (**Fig. 1A**). APA events were widespread and we defined cryptic APA events as ones with <10% mean usage in controls and >10% usage change after TDP-43 knockdown. We identified 227 cryptic APAs to be present in at least 1 dataset (adjusted $P < 0.05$, **Fig. 1B, Supplementary Fig. 1, Supplementary Table 2**). Cryptic ALEs (n=92) included previously identified cryptic exons such as *STMN2*, *ARHGAP32,* and *RSF1* (**Fig. 1B**). 108 3'UTR cryptics were identified, of which 86 are novel 3'UTR extensions (3'Ext; e.g. *TLX1,* **Fig. 1C**) and 20 were 3'UTR shortening events (proximal 3'Ext). 20 IPA events were also detected, including *CNPY3* which was identified and experimentally validated in an orthogonal bioinformatics approach (see Arnold et al., co-submitted). The remaining 9 events could not be uniquely assigned to ALEs or IPAs based on annotation, and are defined as 'complex'.

70% (159/227) of cryptic APAs were detected as significant in a single dataset (**Supplementary Fig. 2A**), but we found that the majority (138) satisfied cryptic expression criteria (<10% mean usage in controls and >10% usage change after TDP-43 knockdown) across datasets. 51 APAs were consistently below 10% usage threshold in controls, but did not sufficiently increase following TDP-43 depletion to meet the cryptic criteria definition across datasets. 28 APAs showed instead a significant increase upon TDP-43 loss across datasets, but had >10% median usage in controls, therefore placing them outside the cryptic criteria, but demonstrating consistent regulation by TDP-43 (**Supplementary Fig. 2B**). Altogether, this data highlights a widespread presence of cryptic APA upon TDP-43 loss.

TDP-43 binding can both repress and enhance polyA site selection

Next, we investigated TDP-43 binding patterns around cryptic APAs using TDP-43 iCLIP data generated in SH-SY5Y cells[10]. We focussed on ALEs and 3'Ext events as the low number of IPA and proximal 3'Ext events (n=20 in both cases) did not allow reliable binding profile inferences. TDP-43 binding was enriched around the splice acceptor of cryptic ALEs as previously described in cryptic splice junctions and downstream of the cryptic polyadenylation site (PAS) of ALEs **(Fig. 1D)**, supporting TDP-43 acting as a repressor of both splicing and polyadenylation. Intriguingly, TDP-43 binding was also enriched immediately downstream of the annotated proximal PAS of 3'Ext events (**Fig. 1E**), supporting a role for TDP-43 in enhancing polyA usage consistent with previous reports of TDP-43 binding with respect to regulated PAS[5].

iCLIP data, typically generated in control cells, is not sensitive in detecting binding to cryptic 3'Ext regions, as these events can be only detected at very low levels with physiological TDP-43 presence. We therefore sought to corroborate our findings by adapting PEKA[24] to infer de-novo hexamer enrichment relative to cryptic landmarks. Previously defined hexamers

enriched around TDP-43 iCLIP binding sites[6] (**Supplementary Fig. 3A**) were overrepresented among the most enriched hexamers proximal to all cryptic landmarks, with the strongest signal overall observed at both the 3'ss and PAS of ALE events (**Supplementary Fig. 3B**). To assess the concordance with iCLIP binding profiles, we visualised the positional coverage of the hexamer group most strongly associated with TDP-43 binding[6]. We observed a striking peak immediately upstream of ALE splice acceptors, consistent with the previously observed mechanism of *STMN2* cryptic exon repression[25] (**Fig. 1D**). Enriched signal was also observed immediately downstream of the distal PAS of 3'Exts (**Fig. 1E)** and the PAS of ALEs (**Fig. 1D**). Overall, our data support a direct role for TDP-43 binding in both enhancing and repressing PAS usage, therefore leading to cryptic APA upon TDP-43 loss.

TDP-43 cryptic APA is detectable in post-mortem ALS/FTD tissues

We next investigated whether the cryptic APA detected *in vitro* occurred also in post-mortem central nervous system (CNS) tissue samples affected by TDP-43 proteinopathy. We initially focused on neuronal nuclei sorted into TDP-43 positive and TDP-43 negative populations[26].

60 cryptic APA events were more highly expressed in TDP-43 depleted nuclei. All APA event types were represented in this list (**Fig. 2A**), with ALEs (28) and 3'Exts (27) representing the majority of enriched events. Our analysis confirms previously reported cryptic ALEs with patient specificity such as in *STMN2*[27]. A number of 3'Ext also show enrichment in TDP-43 negative nuclei in a similar magnitude to *STMN2* (median increased usage of 69 %), most notably *ELK1 (76 %)* and *RBM27 (57 %)* (**Fig. 2A**). Five IPA events meet our enrichment criteria (**Fig. 2A**), including *USP31*, which was identified in a targeted assay of sporadic ALS motor cortex tissue[28]. However, IPA events were generally more weakly enriched in TDP-43 depleted nuclei compared to 3'Ext and ALE events. Altogether, this analysis shows that cryptic APA is detectable in post-mortem ALS/FTD CNS.

Next, we used the New York Genome Centre (NYGC) ALS consortium RNA-seq dataset to assess cryptic APA in a larger cohort of CNS cases with or without TDP-43 pathology. Cryptic 3'Exts often demonstrated low basal expression in control samples in our *in vitro* datasets, confounding the detection in post-mortem bulk RNA-seq datasets, where only a very small proportion of cells is expected to have TDP-43 pathology. IPA detection is further complicated by the fact that normal pre-mRNA reads also map to IPA regions creating significant noise in bulk RNA-seq. We therefore focussed on ALEs, where detection of the associated upstream cryptic splice junctions provide direct evidence of expression. As cryptic ALEs are expected to be dependent on nuclear TDP-43 depletion, we defined criteria based on spliced read detection to identify cryptic events with specific expression in tissues and disease subtypes where TDP-43 pathology is present. 7/118 cryptic ALE junctions fulfilled specificity criteria (**Supplementary Table 3**), in contrast to 56/313 cryptic splicing events collated from i3Neurons with TDP-43 knockdown[13] (**Fig. 2B**). *STMN2* was most frequently detected in tissues with expected TDP-43 proteinopathy, and several other ALEs were amongst the most frequently detected specific cryptic events, including *SYNJ2* (3rd, **Fig. 2C**) and *PHF2* (8th, **Fig. 2D**). Altogether, this suggests that cryptic APAs are detectable in post-mortem tissue affected by TDP-43 pathology, highlighting their potential relevance in loss-of-function disease mechanisms and their promising utility as biomarkers.

Cryptic APA events have variable effects on differential expression

Cryptic splicing events impact expression, often leading to a reduction in transcript levels[9–11]. We therefore assessed the effect of cryptic APAs on their own transcripts in i3Neurons[13] (**Supplementary Fig. 4A**), and found that the majority of events (86/126) coincide with a significant change in expression, equally split between significant upregulation and downregulation. When subdivided further into cryptic APA categories, no category showed a clear bias for upregulation or downregulation (19/34 3'Ext, 17/37 ALE and 6/10 IPA genes are downregulated). This suggests that cryptic APAs are associated with differential expression, but have variable effects on transcript levels.

Cryptic 3'UTR extensions in transcription factor RNAs lead to increased translation and function

Regulation of both ALE and 3'Ext usage has been demonstrated to impact protein abundance through distinct mechanisms[29,30]**,** but differential RNA abundance does not necessarily imply a coordinated change in protein levels. To assess whether changes in gene expression were also reflected in translation levels, we performed differential translation analysis of Ribo-seq data generated from i3Neurons with TDP-43 depletion[13].

Only a minority of cryptic APA-containing genes (26/126) showed significant changes in overall translation levels (**Supplementary Table 4**), of which 24 are concordantly altered in both Ribo-seq and RNA-seq abundance upon TDP-43 KD (**Fig. 3A,3B**). Notably, the differentially translated subset appeared to stratify by APA category: whilst ALEs are downregulated, all four significant 3'Exts, which also showed increased RNA abundance (**Fig. 3A**), had significantly increased translation (**Fig. 3B**). Gene set enrichment analysis (GSEA)[31,32] confirmed that cryptic ALE and 3'Ext genes are significantly associated with decreased (normalised enrichment score (NES) -2.09, padj 2.31e-6) and increased translation (NES 1.54, padj 0.03) respectively, whilst IPA genes show no significant association in either direction (NES -1.09, padj 0.36, **Supplementary Fig. 4B**).

Interestingly, the three 3'Ext-containing genes that were most upregulated at both RNA and translation levels (**Fig. 3A,3B**) encode for three transcription factors (TFs): *ELK1*, *SIX3,* and *TLX1*. The regulation of these 3'Ext events is reproducible across *in vitro* datasets (**Supplementary Fig. 1**). As *ELK1* increase has previously been associated with neuronal toxicity[33–35] and its levels are consistently higher in mature neurons, compared to *SIX3* and *TLX1,* which are associated with neuronal development[36,37], we decided to focus our investigations on *ELK1*. We tested whether the increase in Ribo-seq also corresponded to an upregulation of steady-state protein, and western blots confirmed a significant increase in ELK1 protein expression upon TDP-43 knockdown in i3Neurons (**Fig. 3C**).

We next asked whether the activity of ELK1, which functions as a TF in the ternary complex factor (TCF) family[38], could be altered in the context of TDP-43 loss. We assessed whether ELK1 target genes defined by ChIP-seq in HeLa cells were also affected in TDP-43 knockout HeLa cells[39], where the cryptic 3'Ext is robustly upregulated (**Fig. 3D**). Using GSEA, we observed a significant change in ELK1 target gene expression upon TDP-43 knockout (**Fig. 3D**). This suggests that cryptic 3'Exts can lead to change in function in the context of TDP-43 loss.

TFs with cryptic 3'UTR Extensions have increased RNA stability and cytoplasmic RNA localisation

We investigated the mechanisms by which cryptic 3'UTRs could mediate increased translation levels of *ELK1*, *SIX3*, and *TLX1*. We revisited differential splicing analysis of i3Neuron RNA-seq datasets[10,13] and confirmed that cryptic 3'Exts are the only differential RNA processing events occurring in these 3 TF RNAs upon TDP-43 depletion.

As alternative 3'UTRs have been linked to differences in RNA stability[40], we reasoned that increased RNA stability could account for changes in overall RNA abundance and translation levels. To investigate changes in RNA stability in i3Neurons with TDP-43 depletion, we performed SLAM-seq[41], which allows the detection of newly synthesised RNAs through incorporation of a uridine analogue (4sU). Different lengths of 4sU treatment allow to estimate gene-level RNA half lives. We observed increased half lives in cryptic 3'Ext containing genes *ELK1*, *TLX1,* and *SIX3* (**Fig. 3E**). This suggests that increased RNA abundance and translation of cryptic 3'Ext genes are mediated by increased RNA stability.

Given that translation depends on extra-nuclear localisation of mRNAs, we tested whether altered subcellular localisation of transcripts could be induced by 3'Ext and also contribute to the increased translation levels[42–45]. Focussing on the *ELK1* cryptic 3'Ext, we designed probes to recognise the common proximal sequence and the distal sequence specific to the 3'Ext, and performed fluorescent in-situ hybridisation (FISH) in i3Neurons (**Fig. 3F**). Consistent with RNA-sequencing, upon TDP-43 knockdown we observed a significant increase in total foci for both the total and cryptic-specific probes, with a more pronounced increase for the cryptic-specific probe (**Fig. 3G**). While there was a trend for elevated extranuclear localisation of *ELK1* (**Fig. 3H**), this increase did not reach significance. These findings suggest that increased RNA stability is likely the main driver of increased *ELK1* RNA abundance and translation.

Discussion

Defining TDP-43 RNA targets is critical to understanding the molecular consequences of nuclear TDP-43 depletion. To date, efforts have mainly focussed on the consequences of altered splicing and have successfully identified key targets that are being pursued as therapeutic targets and potential biomarkers for TDP-43 pathology[10,11,14,46,47]. Although TDP-43 is involved in multiple aspects of RNA processing, including polyadenylation[4–6], this has been largely understudied due to the lack of effective tools to address these questions. Here, we develop a pipeline to detect and quantify novel APA events from total RNA-seq and apply it to a wide range of neuronal TDP-43 loss of function datasets to define cryptic APAs, a novel category of cryptic RNA processing events of potential relevance to ALS/FTD. iCLIP and TDP-43 binding motif analyses support a direct regulation of these events by TDP-43, in which TDP-43 loss can both weaken conventional polyA sites and derepress cryptic APA. Similarly to splicing, where TDP-43 can both repress or enhance exon inclusion, TDP-43 can therefore have a dual action on transcript termination. Importantly for disease relevance, and similarly to cryptic splicing, numerous cryptic APA events can be detected in post-mortem tissue and are specifically expressed upon TDP-43 pathology.

When we then moved to investigate the impact of cryptic APAs on RNA levels and translation, and found that IPAs and ALEs had either no impact or induced a reduction of

transcript levels in RNA-seq and Ribo-seq analyses - in line with previous observations on known cryptic ALEs as *STMN2*[46,47]. Recent work has demonstrated that cryptic exon-containing transcripts can be translated and produce cryptic peptides that could serve as biomarkers of TDP-43 pathology[13,14]. As cryptic ALE and IPA events are mostly predicted to be insensitive to nonsense mediated decay, they are likely to give rise to cryptic peptides; e.g. cryptic ALE *RSF1* encodes a cryptic peptide that is detected in CSF of ALS patients[13]. Previous work has identified cryptic ALEs as their novel splice junction can be detected by numerous splice-detection packages[13,27,48]. Conversely, IPAs have been harder to identify and further work should consider whether these cryptic IPA events can be detected in patient brains and biofluids as an indirect measure of TDP-43 pathology.

Surprisingly 3'Ext events in the three transcription factors ELK1, SIX3, and TLX1 were associated with transcript upregulation, increased translation and protein levels. We found this to be associated with an increase in RNA stability leading to an increase in cytoplasmic localisation. Thus, in contrast to the conventional model of TDP-43 regulated cryptic splicing leading to reduced protein levels or to altered proteins containing cryptic peptides, cryptic 3'Ext can be associated with increased protein levels, outlining a novel consequence of TDP-43 cryptic RNA processing.

ELK1, *SIX3,* and *TLX1* 3'Ext are reliably induced upon TDP-43 depletion across our *in vitro* datasets, suggesting they are not cell-type specific, sensitive TDP-43 targets. These three TFs have been studied in the neuronal context, although *SIX3* and *TLX1* are primarily expressed in the developmental stage[36,37]. Our work therefore focused on *ELK1* and we were able to utilise HeLa cell data in order to show that TDP-43 loss can induce changes in *ELK1* target genes. *ELK1* promotes axonal outgrowth[49] and is increased in Huntington's disease models where it can have a neuroprotective role[50]. *ELK1* overexpression has also been linked with neurotoxicity through interaction with components of the mitochondrial permeability-transition pore complex[34], and dendrite-specific overexpression of *ELK1* mRNA induced cell death in a transcription- and translation-dependent manner[33], supporting a potential contribution of this cryptic APA to pathogenesis. Further work is needed to investigate the functional relevance of increased *ELK1*, *SIX3*, and *TLX1* expression in models of TDP-43 proteinopathy.

We focussed on identifying cryptic APA events, as their extreme expression changes upon TDP-43 loss renders them favourable therapeutic and biomarker targets. Accompanying manuscripts by Arnold et al. and Zeng et al. investigate APA dysregulation more generally upon TDP-43 loss and show it is widespread, in accordance with our findings in **Fig. 1B**, can occur in ALS-FTD related genes (Zeng et al., co-submitted) and can lead to change in function (Arnold et al., co-submitted), underscoring the potential relevance of APA in disease pathogenesis. We note that several targets (e.g. *CNPY3, ELK1, ARHGAP32)* are commonly identified across the studies despite diverging methodological approaches, underlying the consistency of our observations. Importantly, similarly to our findings for *ELK1*, *SIX3*, and *TLX1*, both Zeng et al. and Arnold et al. manuscripts also find that APAs can lead to upregulation of normal protein levels, consolidating this a general consequence of TDP-43 loss. Our studies collectively demonstrate that dysregulated APA is a general consequence of nuclear TDP-43 loss in ALS-FTD.

In summary, we provide a compendium of cryptic APA events determined by TDP-43 loss as a resource for studying RNA dysregulation and identifying novel biomarkers in ALS. Our work also shows that cryptic RNA processing can lead to an increase in protein expression and function, expanding the molecular consequences of TDP-43 loss and pathology, with implications for disease pathogenesis and therapeutic target identification.

References

1. Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M., et al. (2006). Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. Science *314*, 130–133. 10.1126/science.1134108.
2. Neumann, M., Tolnay, M., and Mackenzie, I.R.A. (2009). The molecular basis of frontotemporal dementia. Expert Rev. Mol. Med. *11*, e23. 10.1017/S1462399409001136.
3. Meneses, A., Koga, S., O'Leary, J., Dickson, D.W., Bu, G., and Zhao, N. (2021). TDP-43 Pathology in Alzheimer's Disease. Mol. Neurodegener. *16*, 84. 10.1186/s13024-021-00503-x.
4. Eréndira Avendaño-Vázquez, S., Dhir, A., Bembich, S., Buratti, E., Proudfoot, N., and Baralle, F.E. (2012). Autoregulation of TDP-43 mRNA levels involves interplay between transcription, splicing, and alternative polyA site selection. Genes Dev. *26*, 1679–1684. 10.1101/gad.194829.112.
5. Rot, G., Wang, Z., Huppertz, I., Modic, M., Lenče, T., Hallegger, M., Haberman, N., Curk, T., von Mering, C., and Ule, J. (2017). High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. Cell Rep. *19*, 1056–1067. 10.1016/j.celrep.2017.04.028.
6. Hallegger, M., Chakrabarti, A.M., Lee, F.C.Y., Lee, B.L., Amalietti, A.G., Odeh, H.M., Copley, K.E., Rubien, J.D., Portz, B., Kuret, K., et al. (2021). TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. Cell *184*, 4680-4696.e22. 10.1016/j.cell.2021.07.018.
7. Ratti, A., and Buratti, E. (2016). Physiological functions and pathobiology of TDP-43 and FUS/TLS proteins. J. Neurochem. *138*, 95–111. 10.1111/jnc.13625.
8. Mehta, P.R., Brown, A.-L., Ward, M.E., and Fratta, P. (2023). The era of cryptic exons: implications for ALS-FTD. Mol. Neurodegener. *18*, 16. 10.1186/s13024-023-00608-5.
9. Ling, J.P., Pletnikova, O., Troncoso, J.C., and Wong, P.C. (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science *349*, 650–655. 10.1126/science.aab0983.
10. Brown, A.-L., Wilkins, O.G., Keuss, M.J., Hill, S.E., Zanovello, M., Lee, W.C., Bampton, A., Lee, F.C.Y., Masino, L., Qi, Y.A., et al. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. Nature *603*, 131–137. 10.1038/s41586-022-04436-3.
11. Ma, X.R., Prudencio, M., Koike, Y., Vatsavayai, S.C., Kim, G., Harbinski, F., Briner, A., Rodriguez, C.M., Guo, C., Akiyama, T., et al. (2022). TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. Nature *603*, 124–130. 10.1038/s41586-022-04424-7.
12. Humphrey, J., Emmett, W., Fratta, P., Isaacs, A.M., and Plagnol, V. (2017). Quantitative analysis of cryptic splicing associated with TDP-43 depletion. BMC Med. Genomics *10*, 38. 10.1186/s12920-017-0274-1.

13. Seddighi, S., Qi, Y.A., Brown, A.-L., Wilkins, O.G., Bereda, C., Belair, C., Zhang, Y., Prudencio, M., Keuss, M.J., Khandeshi, A., et al. (2023). Mis-spliced transcripts generate de novo proteins in TDP-43-related ALS/FTD. Preprint at bioRxiv, 10.1101/2023.01.23.525149 10.1101/2023.01.23.525149.
14. Irwin, K.E., Jasin, P., Braunstein, K.E., Sinha, I., Bowden, K.D., Moghekar, A., Oh, E.S., Raitcheva, D., Bartlett, D., Berry, J.D., et al. (2023). A fluid biomarker reveals loss of TDP-43 splicing repression in pre-symptomatic ALS. Preprint at bioRxiv, 10.1101/2023.01.23.525202 10.1101/2023.01.23.525202.
15. Neve, J., Patel, R., Wang, Z., Louey, A., and Furger, A.M. (2017). Cleavage and polyadenylation: Ending the message expands gene regulation. RNA Biol. *14*, 865–890. 10.1080/15476286.2017.1306171.
16. Mitschka, S., and Mayr, C. (2022). Context-specific regulation and function of mRNA alternative polyadenylation. Nat. Rev. Mol. Cell Biol., 1–18. 10.1038/s41580-022-00507-5.
17. LaForce, G.R., Farr, J.S., Liu, J., Akesson, C., Gumus, E., Pinkard, O., Miranda, H.C., Johnson, K., Sweet, T.J., Ji, P., et al. (2022). Suppression of premature transcription termination leads to reduced mRNA isoform diversity and neurodegeneration. Neuron. 10.1016/j.neuron.2022.01.018.
18. Singh, I., Lee, S.H., Sperling, A.S., Samur, M.K., Tai, Y.T., Fulciniti, M., Munshi, N.C., Mayr, C., and Leslie, C.S. (2018). Widespread intronic polyadenylation diversifies immune cell transcriptomes. Nat. Commun. *9*, 1–16. 10.1038/s41467-018-04112-z.
19. Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. *20*, 278. 10.1186/s13059-019-1910-1.
20. Herrmann, C.J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A.J., and Zavolan, M. (2019). PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. Nucleic Acids Res. 10.1093/nar/gkz918.
21. Gruber, A.J., Schmidt, R., Gruber, A.R., Martin, G., Ghosh, S., Belmadani, M., Keller, W., and Zavolan, M. (2016). A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. Genome Res. *26*, 1145–1159. 10.1101/gr.202432.115.
22. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. 10.1038/nmeth.4197.
23. Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017. 10.1101/gr.133744.111.
24. Kuret, K., Amalietti, A.G., Jones, D.M., Capitanchik, C., and Ule, J. (2022). Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP. Genome Biol. *23*, 191. 10.1186/s13059-022-02755-2.
25. Baughn, M.W., Melamed, Z., López-Erauskin, J., Beccari, M.S., Ling, K., Zuberi, A., Presa, M., Gonzalo-Gil, E., Maimon, R., Vazquez-Sanchez, S., et al. (2023). Mechanism of STMN2 cryptic splice-polyadenylation and its correction for TDP-43 proteinopathies. Science *379*, 1140–1149. 10.1126/science.abq5622.
26. Liu, E.Y., Russ, J., Cali, C.P., Phan, J.M., Amlie-Wolf, A., and Lee, E.B. (2019). Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. Cell Rep. *27*, 1409-1421.e6. 10.1016/j.celrep.2019.04.003.
27. Prudencio, M., Humphrey, J., Pickles, S., Brown, A.-L., Hill, S.E., Kachergus, J.M., Shi, J., Heckman, M.G., Spiegel, M.R., Cook, C., et al. (2020). Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. J. Clin. Invest. *130*. 10.1172/JCI139741.
28. Cao, M.C., Ryan, B., Wu, J., Curtis, M.A., Faull, R.L.M., Dragunow, M., and Scotter, E.L. (2023). A panel of TDP-43-regulated splicing events verifies loss of TDP-43 function in amyotrophic lateral sclerosis brain tissue. Neurobiol. Dis., 106245.

10.1016/j.nbd.2023.106245.
29. de Prisco, N., Ford, C., Elrod, N.D., Lee, W., Tang, L.C., Huang, K.-L., Lin, A., Ji, P., Jonnakuti, V.S., Boyle, L., et al. (2023). Alternative polyadenylation alters protein dosage by switching between intronic and 3′UTR sites. Sci. Adv. *9*, eade4814. 10.1126/sciadv.ade4814.
30. Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. eLife *5*, e10921. 10.7554/eLife.10921.
31. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550. 10.1073/pnas.0506580102.
32. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. Preprint at bioRxiv, 10.1101/060012 10.1101/060012.
33. Barrett, L.E., Sul, J.-Y., Takano, H., Van Bockstaele, E.J., Haydon, P.G., and Eberwine, J.H. (2006). Region-directed phototransfection reveals the functional significance of a dendritically synthesized transcription factor. Nat. Methods *3*, 455–460. 10.1038/nmeth885.
34. Barrett, L.E., Bockstaele, E.J.V., Sul, J.Y., Takano, H., Haydon, P.G., and Eberwine, J.H. (2006). Elk-1 associates with the mitochondrial permeability transition pore complex in neurons. Proc. Natl. Acad. Sci. *103*, 5155–5160. 10.1073/pnas.0510477103.
35. Sharma, A., Callahan, L.M., Sul, J.-Y., Kim, T.K., Barrett, L., Kim, M., Powers, J.M., Federoff, H., and Eberwine, J. (2010). A Neurotoxic Phosphoform of Elk-1 Associates with Inclusions from Multiple Neurodegenerative Diseases. PLOS ONE *5*, e9002. 10.1371/journal.pone.0009002.
36. Kumar, J.P. (2009). The sine oculis homeobox (SIX) family of transcription factors as regulators of development and disease. Cell. Mol. Life Sci. CMLS *66*, 565–583. 10.1007/s00018-008-8335-4.
37. Cheng, L., Arata, A., Mizuguchi, R., Qian, Y., Karunaratne, A., Gray, P.A., Arata, S., Shirasawa, S., Bouchard, M., Luo, P., et al. (2004). Tlx3 and Tlx1 are post-mitotic selector genes determining glutamatergic over GABAergic cell fates. Nat. Neurosci. *7*, 510–517. 10.1038/nn1221.
38. Besnard, A., Galan, B., Vanhoutte, P., and Caboche, J. (2011). Elk-1 a Transcription Factor with Multiple Facets in the Brain. Front. Neurosci. *5*, 35. 10.3389/fnins.2011.00035.
39. Roczniak-Ferguson, A., and Ferguson, S.M. (2019). Pleiotropic requirements for human TDP-43 in the regulation of cell and organelle homeostasis. Life Sci. Alliance *2*. 10.26508/lsa.201900358.
40. Zheng, D., Wang, R., Ding, Q., Wang, T., Xie, B., Wei, L., Zhong, Z., and Tian, B. (2018). Cellular stress alters 3′UTR landscape through alternative polyadenylation and isoform-specific degradation. Nat. Commun. *9*. 10.1038/s41467-018-04730-7.
41. Herzog, V.A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T.R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S.L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. Nat. Methods *14*, 1198–1204. 10.1038/nmeth.4435.
42. Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M., and Schuman, E.M. (2018). Alternative 3′ UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron *98*, 495-511.e6. 10.1016/j.neuron.2018.03.030.
43. Taliaferro, J.M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E.T., Graveley, B.R., Gertler, F.B., Swanson, M.S., et al. (2016). Distal Alternative Last Exons Localize mRNAs to Neural Projections. Mol. Cell *61*, 821–833. 10.1016/j.molcel.2016.01.020.
44. Arora, A., Goering, R., Lo, H.Y.G., Lo, J., Moffatt, C., and Taliaferro, J.M. (2022). The Role of Alternative Polyadenylation in the Regulation of Subcellular RNA Localization. Front.

Genet. *12*.
45. Mattioli, C.C., Rom, A., Franke, V., Imami, K., Arrey, G., Terne, M., Woehler, A., Akalin, A., Ulitsky, I., and Chekulaeva, M. (2019). Alternative 3 UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Res. *47*, 2560–2573. 10.1093/nar/gky1270.
46. Melamed, Z., López-Erauskin, J., Baughn, M.W., Zhang, O., Drenner, K., Sun, Y., Freyermuth, F., McMahon, M.A., Beccari, M.S., Artates, J.W., et al. (2019). Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. Nat. Neurosci. *22*, 180–190. 10.1038/s41593-018-0293-z.
47. Klim, J.R., Williams, L.A., Limone, F., Guerra San Juan, I., Davis-Dusenbery, B.N., Mordes, D.A., Burberry, A., Steinbaugh, M.J., Gamage, K.K., Kirchner, R., et al. (2019). ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. Nat. Neurosci. *22*, 167–179. 10.1038/s41593-018-0300-4.
48. Gittings, L.M., Alsop, E.B., Antone, J., Singer, M., Whitsett, T.G., Sattler, R., and Van Keuren-Jensen, K. (2023). Cryptic exon detection and transcriptomic changes revealed in single-nuclei RNA sequencing of C9ORF72 patients spanning the ALS-FTD spectrum. Acta Neuropathol. (Berl.). 10.1007/s00401-023-02599-5.
49. Noro, T., Shah, S.H., Yin, Y., Kawaguchi, R., Yokota, S., Chang, K.-C., Madaan, A., Sun, C., Coppola, G., Geschwind, D., et al. (2022). Elk-1 regulates retinal ganglion cell axon regeneration after injury. Sci. Rep. *12*, 17446. 10.1038/s41598-022-21767-3.
50. Anglada-Huguet, M., Giralt, A., Perez-Navarro, E., Alberch, J., and Xifró, X. (2012). Activation of Elk-1 participates as a neuroprotective compensatory mechanism in models of Huntington's disease. J. Neurochem. *121*, 639–648. 10.1111/j.1471-4159.2012.07711.x.
51. Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron *104*, 239-255.e12. 10.1016/j.neuron.2019.07.014.
52. Fernandopulle, M.S., Prestil, R., Grunseich, C., Wang, C., Gan, L., and Ward, M.E. (2018). Transcription Factor–Mediated Differentiation of Human iPSCs into Neurons. Curr. Protoc. Cell Biol. *79*, e51. 10.1002/cpcb.51.
53. Nehme, R., Zuccaro, E., Ghosh, S.D., Li, C., Sherwood, J.L., Pietilainen, O., Barrett, L.E., Limone, F., Worringer, K.A., Kommineni, S., et al. (2018). Combining NGN2 Programming with Developmental Patterning Generates Human Excitatory Neurons with NMDAR-Mediated Synaptic Transmission. Cell Rep. *23*, 2509–2523. 10.1016/j.celrep.2018.04.066.
54. Bardy, C., van den Hurk, M., Eames, T., Marchand, C., Hernandez, R.V., Kellogg, M., Gorris, M., Galet, B., Palomares, V., Brown, J., et al. (2015). Neuronal medium that supports basic synaptic functions and activity of human neurons in vitro. Proc. Natl. Acad. Sci. *112*, E2725–E2734. 10.1073/pnas.1504393112.
55. Tovell, H., Testa, A., Maniaci, C., Zhou, H., Prescott, A.R., Macartney, T., Ciulli, A., and Alessi, D.R. (2019). Rapid and Reversible Knockdown of Endogenously Tagged Endosomal Proteins via an Optimized HaloPROTAC Degrader. ACS Chem. Biol. *14*, 882–892. 10.1021/acschembio.8b01016.
56. Humphrey, J., Venkatesh, S., Hasan, R., Herb, J.T., de Paiva Lopes, K., Küçükali, F., Byrska-Bishop, M., Evani, U.S., Narzisi, G., Fagegaltier, D., et al. (2023). Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. Nat. Neurosci. *26*, 150–162. 10.1038/s41593-022-01205-3.
57. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. 10.1093/bioinformatics/btu170.
58. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C.,

Wright, J.C., Armstrong, J., Barnes, I., et al. (2020). GENCODE 2021. Nucleic Acids Res. *49*, D916–D923. 10.1093/nar/gkaa1087.
59. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.
60. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. 10.1093/bioinformatics/bty560.
61. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. Preprint at F1000Research, 10.12688/f1000research.29032.2 10.12688/f1000research.29032.2.
62. Rummel, T., Sakellaridi, L., and Erhard, F. (2023). grandR: a comprehensive package for nucleotide conversion RNA-seq data analysis. Nat. Commun. *14*, 3559. 10.1038/s41467-023-39163-4.
63. Shah, A., Mittleman, B.E., Gilad, Y., and Li, Y.I. (2021). Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. Genome Biol. *22*, 291. 10.1186/s13059-021-02502-z.
64. Bryce-Smith, S., Burri, D., Gazzara, M.R., Herrmann, C.J., Danecka, W., Fitzsimmons, C.M., Wan, Y.K., Zhuang, F., Fansler, M.M., Fernández, J.M., et al. (2023). Extensible benchmarking of methods that identify and quantify polyadenylation sites from RNA-seq data. RNA *29*, 1839–1855. 10.1261/rna.079849.123.
65. Lusk, R., Stene, E., Banaei-Kashani, F., Tabakoff, B., Kechris, K., and Saba, L.M. (2021). Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. Nat. Commun. *12*, 1652. 10.1038/s41467-021-21894-x.
66. Gruber, A.J., Gypas, F., Riba, A., Schmidt, R., and Zavolan, M. (2018). Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. Nat. Methods *15*, 832–836. 10.1038/s41592-018-0114-z.
67. Li, W.V., Li, S., Tong, X., Deng, L., Shi, H., and Li, J.J. (2019). AIDE: annotation-assisted isoform discovery with high precision. Genome Res. *29*, 2056–2072. 10.1101/gr.251108.119.
68. Sethi, S., Zhang, D., Guelfi, S., Chen, Z., Garcia-Ruiz, S., Olagbaju, E.O., Ryten, M., Saini, H., and Botia, J.A. (2022). Leveraging omic features with F3UTER enables identification of unannotated 3'UTRs for synaptic genes. Nat. Commun. *13*, 2270. 10.1038/s41467-022-30017-z.
69. Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. Preprint at F1000Research, 10.12688/f1000research.23297.2 10.12688/f1000research.23297.2.
70. Swamy, V.S., Fufa, T.D., Hufnagel, R.B., and McGaughey, D.M. (2020). A long read optimized de novo transcriptome pipeline reveals novel ocular developmentally regulated gene isoforms and disease targets. Preprint at bioRxiv, 10.1101/2020.08.21.261644 10.1101/2020.08.21.261644.
71. Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M.I., Kingsford, C., and Patro, R. (2020). Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. *21*, 239. 10.1186/s13059-020-02151-8.
72. Soneson, C., Love, M.I., and Robinson, M.D. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. F1000Research *4*, 1521. 10.12688/F1000RESEARCH.7563.2.
73. Stovner, E.B., and Sætrom, P. (2020). PyRanges: efficient comparison of genomic intervals in Python. Bioinformatics *36*, 918–919. 10.1093/bioinformatics/btz615.
74. Shirley, M.D., Ma, Z., Pedersen, B.S., and Wheelan, S.J. (2015). Efficient "pythonic" access to FASTA files using pyfaidx (PeerJ Inc.) 10.7287/peerj.preprints.970v1.

75. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422. 10.1093/bioinformatics/btp163.
76. Goering, R., Engel, K.L., Gillen, A.E., Fong, N., Bentley, D.L., and Taliaferro, J.M. (2021). LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival. BMC Genomics *22*, 476. 10.1186/s12864-021-07781-1.
77. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.
78. Amalietti, A.G. (2021). Comparative Visualisation of Average Motif Coverage. Version 1.1.0. 10.5281/zenodo.8386510 10.5281/zenodo.8386510.
79. Tam, O.H., Rozhkov, N.V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., Propp, N., Phatnani, H., Kwan, J., Sareen, D., et al. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. Cell Rep. *29*, 1164-1177.e5. 10.1016/j.celrep.2019.09.066.
80. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. 10.1093/bioinformatics/btt656.
81. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. 10.1186/s13059-014-0550-8.
82. Zhu, A., Ibrahim, J.G., and Love, M.I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics *35*, 2084–2092. 10.1093/bioinformatics/bty895.
83. Zou, Z., Ohta, T., Miura, F., and Oki, S. (2022). ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res. *50*, W175–W182. 10.1093/nar/gkac199.
84. Boros, J., O'Donnell, A., Donaldson, I.J., Kasza, A., Zeef, L., and Sharrocks, A.D. (2009). Overlapping promoter targeting by Elk-1 and other divergent ETS-domain transcription factor family members. Nucleic Acids Res. *37*, 7368–7380. 10.1093/nar/gkp804.
85. Vaquero-Garcia, J., Aicher, J.K., Jewell, S., Gazzara, M.R., Radens, C.M., Jha, A., Norton, S.S., Lahens, N.F., Grant, G.R., and Barash, Y. (2023). RNA splicing analysis using heterogeneous and large RNA-seq datasets. Nat. Commun. *14*, 1230. 10.1038/s41467-023-36585-y.
86. R Core Team (2023). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
87. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).
88. Kassambara, A. (2023). ggpubr: "ggplot2" Based Publication Ready Plots.
89. Dawson, C. (2022). ggprism: A "ggplot2" Extension Inspired by "GraphPad Prism" 10.5281/zenodo.4556067.
90. Slowikowski, K. (2024). ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2."
91. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686. 10.21105/joss.01686.
92. Ooms, J. (2024). writexl: Export Data Frames to Excel "xlsx" Format. Version 1.4.2.
93. Barrett, T., Dowle, M., and Srinivasan, A. (2023). data.table: Extension of `data.frame`.
94. The pandas development (2023). pandas-dev/pandas: Pandas. Version 2.0.2 (Zenodo). 10.5281/zenodo.3509134 10.5281/zenodo.3509134.

95. Harris, C.R., Millman, K.J., Walt, S.J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. 10.1038/s41586-020-2649-2.
96. conda contributors (2024). conda: A system-level, binary package and environment manager running on all major operating systems and platforms.
97. Ushey, K., and Wickham, H. (2023). renv: Project Environments.

Methods

CRISPRi knockdown in human iPS cells, and differentiation and culture of i3Neurons

CRISPRi knockdown experiments were performed in the WTC11 iPSC line harbouring stable TO-NGN2 and dCas9-BFP-KRAB cassettes at safe harbour loci[51]. CRISPRi knockdown of TDP-43 in iPSCs was achieved using sgRNA targeting the transcription start site of TARDBP (or non-targeting control sgRNA)[10], delivered by lentiviral transduction. sgRNA sequences were as follows: non-targeting control GTCCACCCTTATCTAGGCTA; and TARDBP GGGAAGTCAGCCGTGAGACC. iPS cells were differentiated into cortical-like i3Neurons as described previously[10,52] and fixed 9 days after re-plating for RNA-fluorescence in-situ hybridisation (FISH). For RNA-seq experiments ('Humphrey i3 cortical'), i3Neurons were induced as previously described[52] with the addition of SMAD and WNT inhibitors[53] (SB431542 10 µM; LDN-193189 100 nM; XAV939 2 µM all from Cambridge Biosciences). After induction, cells were cultured in BrainPhys Media (StemCell Technologies) with 20 ng/ml BDNF (PeproTech), 20 ng/ml GDNF (PeproTech), 1x N2 supplement (Thermo Fisher), 1 x B27 supplement (Thermo Fisher), 200 nM Ascorbic Acid (Sigma), 1 mM dibutyryl cyclic-AMP (Sigma) and 1 µg/ml Laminin (Thermo Fisher) as previously described[54] and harvested 30 days after differentiation.

An iPSC line with a N-terminal HaloTag on both endogenous copies of TDP-43 (Halo-TDP-43 i3Neurons) was generated by CRISPR/Cas12 gene editing. The parental cell line used was the WTC11 cell line with integrated dCas9-Krab and NGN2 cassettes as mentioned previously[51]. The HDR template used was addgene plasmid 178131. Editing was done with Cas12 crRNA (Integrated DNA Technologies) with GGAAAAGTAAAAGATGTCTGAAT as the targeting sequence. Recombinant Cas12 (Cpf1 ultra, Integrated DNA Technologies) was electroporated with HDR template and Cas12 crRNA using the P3 Primary Cell 4-D Nucleofector kit (V4XP-3024 Amaxa). iPSCs were then single cell plated and positive colonies were selected with HaloTag TMR dye (Promega) and verified by PCR of genomic DNA.

For PROTAC mediated knockdown of Halo-TDP-43, i3Neurons were treated with HaloPROTAC-E[55] (30 nM) on DIV14 and harvested on DIV28.

Strand-specific, poly(A) enriched sequencing libraries for the 'Humphrey i3 cortical' dataset were prepared using the Kapa mRNA Hyper Prep kit. 100ng total RNA was used as input material for poly(A)+ mRNA capture. Fragmentation was performed for 6 minutes at 85°C to obtain a target fragment size of 300-400bp, and 13 cycles of PCR amplification were performed. The resulting libraries were sequenced 2x150bp on the Illumina NextSeq2000 machine. Samples were

processed as previously described[56] using the RAPiD-nf nextflow pipeline. Briefly, adapters were trimmed from raw reads using Trimmomatic[57] v0.36, and reads were aligned to the GRCh38 genome build using gene models from GENCODE v30[58] with STAR[59] v2.7.2a. The RAPiD-nf pipeline is available at https://github.com/CommonMindConsortium/RAPiD-nf/ .

Fluorescent in-situ hybridisation (FISH)

Cortical-like i3Neurons were cultured on 13-mm glass coverslips and fixed in 4% PFA/sucrose on day 9. RNA-FISH was performed using the QuantiGene ViewRNA ISH Cell Assay kit (Invitrogen, QVC0001), according to manufacturer's instructions. Protease was used at 1:1000 dilution. Two probe sets were used to detect the canonical *ELK1* transcript (TYPE 4 probe, 488-nm) or specifically the distal 3'UTR cryptic extension (TYPE 1 probe, 550-nm). Confocal images were acquired with a LSM980 laser-scanning confocal microscope with Airyscan 2 (Zeiss), using a 40X oil immersion objective.

For each biological replicate, ten images were acquired for the control and TDP-43 knockdown conditions. For each image, foci for both probes were counted within the 106.07 microns by 106.07 microns field of view on Fiji/ImageJ using the maximum intensity Z-projection function to flatten the 2-µm-thick Z-stack. The Find Maxima function using the same prominence setting between conditions was performed to quantify total numbers of RNA foci. To separately count nuclear and cytoplasmic foci, the Cell Counter plugin was used. For each probe and field of view, the total number of foci was divided by the number of DAPI-stained nuclei to give the average number of foci/cell. To calculate the nuclear:extra-nuclear ratio for the 'Total *ELK1*' probe, the number of nuclear foci was divided by the number of extra-nuclear foci in each field of view. For each probe and condition, the mean number of foci/cell and nuclear:extra-nuclear ratio was calculated from the ten images and normalised, for each biological replicate, to the respective control condition. Statistical significance was evaluated using a one sample t-test with a logarithmic transformation and the Benjamini and Hochberg false discovery rate procedure, testing the null hypothesis that mean = log(1).

Western blots

Halo-TDP-43 i3Neurons were homogenised in lysis buffer (25 mM Tris-HCl, 150 mM NaCl, 1% NP-40, 1% Glycerol, 2 mM EDTA, 0,1% SDS, protease inhibitor (cOmplete™ EDTA-free protease inhibitor cocktail, Roche), phosphatase inhibitor (PhoSTOP™, Roche)). Samples were loaded on a NuPAGE 4-12% Bis-Tris protein gel (Invitrogen), which was run in NuPAGE MOPS buffer. Proteins were transferred onto PVDF blotting membrane (Amersham), through wet transfer for 1h 30 min at 200 mA in transfer buffer (25 mM Tris, 192 mM glycine, 20% methanol). The membrane was blocked in 5% milk in TBST (20 mM Tris, 150 mM NaCl, 0.1% Tween-20) and incubated overnight with primary antibodies diluted in 5% milk in TBST (anti-ELK1 (Abcam ab32106) 1:500, anti-TDP-43 (Abcam, ab104223) 1:2000, anti-tubulin (Sigma-Aldrich, MAB1637) 1:5000). After 1h-incubation with HRP-conjugated secondary antibodies diluted in 5% milk in TBST (anti-mouse HRP (BioRad, 1706516) 1:10000, anti-rabbit HRP (BioRad 1706515) 1:10000), the membrane was developed using Immobilon Classico HRP substrate (Sigma) and the Bio-Rad ChemiDoc system.

SH-SY5Y & SK-N-BE(2) TDP-43 KD and sequencing

SH-SY5Y and SK-N-BE(2) cells were transduced with a SmartVector lentivirus (V3IHSHEG_6494503) containing a doxycycline-inducible shRNA cassette for TDP-43. Transduced cells were selected with puromycin (1 µg/ml) for one week, before being plated as single cells and expanded to obtain a clonal population. Cells were grown in DMEM/F12 + Glutamax (Thermo) supplemented with 10% FBS (Thermo) and 1% PenStrep (Thermo).

For induction of shRNA against TDP-43 cells were treated with the following amounts of doxycycline hyclate (Sigma), and collected after 10 days:
- For experiments in SH-SY5Y cells (curves), 75 ng/ml.
- For experiments in SH-SY5Y cells (CHX), 25 ng/ml.
- For experiments in SK-N-BE(2) cells, 1000 ng/ml.

RNA was extracted from SH-SY5Y and SK-N-BE(2) cells using the RNeasy mini kit (Qiagen) following the manufacturer's protocol including the on-column DNA digestion step. RNA concentrations were measured by Nanodrop and 1,000 ng of RNA was used for reverse transcription. Samples undergoing RNA sequencing were furthermore assessed for RNA quality on a TapeStation 4200 (Agilent), resulting in RNA integrity number (RIN) above 9.4 for all samples.

Sequencing libraries were prepared with polyA enrichment using a TruSeq Stranded mRNA Prep Kit (Illumina) and sequenced on an Illumina HiSeq 2500 or NovaSeq 6000 machine at UCL Genomics with the following specifics:
- SH-SY5Y cells: 2×100 bp, depth > 40M/sample
- SK-N-BE(2) cells: 2x150 bp, depth > 40M/sample

RNA-seq data processing

The 'Brown' SH-SY-5Y, SK-N-BE(2) and i3Neuron datasets were processed as previously described[10]. Unless otherwise stated, all short-read RNA-seq datasets were processed using the following pipeline. Raw reads in FASTQ format were quality trimmed for a minimum Phred score of 10 and otherwise default parameters using fastp[60] (v0.20.1). Quality trimmed reads were aligned to the GRCh38 genome build using gene models from GENCODE v40[58] with STAR[59] (v2.7.8a). Quality trimmed reads are used as input for any tools that require FASTQ files as input (e.g. PAPA, Salmon). Our alignment pipeline is implemented in Snakemake[61] and is available at https://github.com/frattalab/rna_seq_snakemake.

SLAM-seq

SLAM-sequencing was performed on cortical-like i3Neurons following protocols adapted from Herzog et al.[41]. Samples were treated with 100 µM 4sU on Day 7 for 0, 1, 4, 8, 12, 24 hours before immediate wash with phosphate buffered saline (PBS). Each time point had 2 replicates

for both control and TDP-43 knockdown excluding 4 hours where one of the control replicates did not pass RNA quality controls and so was not submitted for sequencing.

RNA was extracted using the Qiagen RNA isolation and purification kit. RNA concentration was estimated using a Nanodrop Microvolume Spectrophotometer (Thermofisher). After ensuring an adequate amount of RNA in each sample, iodoacetamide (IAA) treatment was applied to each, facilitating the thiol modification of incorporated 4sU.

Sequencing libraries were prepared with Kapa RiboErase RNA Hyper kit and sequenced (2 × 250 bp) on an Illumina NovaSeq SP. Using the 'rna_seq_snakemake' alignment pipeline (https://github.com/frattalab/rna_seq_snakemake), raw FASTQ files were quality trimmed using fastp[60] with the parameter "qualified_quality_phred: 10", and aligned without soft-clipping to the GRCh38 genome build using STAR[59] (v2.7.0f) with gene models from GENCODE v34[58]. GRAND-SLAM (v2.0.7b) was run on the aligned data using gene models from GENCODE v34[58] using the "-trim5p 10 -trim3p 10" parameter to ignore mismatches at the ends of reads. The output files containing the estimated new-to-total RNA ratios (NTR) of each gene were used to estimate the half-life of each gene using the recommended workflow in grandR[62] .

PAPA - pipeline to detect cryptic last exons

Whilst there are many tools for de-novo detection of alternative polyadenylation events within 3'UTRs events from RNA-seq data, all of them suffer from poor performance with respect to matched 3'end sequencing approaches[63,64]. Few tools have focused on the detection of gene-body internal poly(A) sites and definition of the complete last exon structure, which is essential if one aims to predict putative encoded peptides. Aptardi is a deep-learning based approach to refine predicted 3'ends of reference or assembled transcriptomes[65], but a benchmarking study found the compute times and resources to be unviable for performance evaluation[64]. TECTool trains a machine-learning model on annotated last exons using transcriptomic features to classify novel intronic last exons defined upstream of polyA sites from the PolyASite atlas[66]. However, as of v0.4 it can only define ALEs and only supports single-end RNA-seq data, which would substantially impact sensitivity to detect events.

General purpose bulk RNA-seq transcript assemblers such as StringTie[19] could be used to identify all alternative novel last exon structures, but they show poor performance in defining precise terminal exon boundaries[66]. Previous benchmarking attempts have suggested that transcript assemblers exhibit poor performance at defining full transcript structures, but much better performance at defining individual exons[67]. This suggests that a scalable solution could be to refine the 3'end predictions of exons predicted by transcript assemblers.

Broadly inspired by a previous workflow combining matched short read and 3' enriched sequencing[18], our approach is to extract last exons from StringTie assembled transcripts and filter based on proximity to 3'end-seq derived PolyA site annotations and/or presence of polyA signal sequences in the terminal predicted region. PolyA signal hexamer presence previously

emerged as one of the most important features in discriminating intergenic expressed regions as 3'UTRs from other transcriptomic regions[68].

Pipeline setup

Transcript assemblies for individual samples are generated using StringTie v 2.1.7 (default settings) in annotation-guided mode. Individual sample assemblies are grouped according to their experimental condition and merged into a redundant assembly using Gffcompare[69] v0.11.2. Condition-wise mean TPMs are calculated for each transcript, assigning a TPM of 0 if a transcript was not assembled in a particular sample. Transcripts are subsequently filtered for a minimum mean expression > 1 TPM, on the basis of a previous study which proposed such a filter to improve global accuracy of assembled transcripts with respect to matched long-read sequencing[70].

Last exons are extracted from expression-filtered transcripts in a sample-wise manner using a custom script. Novel last exons are extracted using the following criteria:

- All predicted 3'ends must not overlap with any annotated exon
- The last intron of ALEs must be contained within an annotated intron and its 5'ss must exactly match an annotated 5'ss
- The last intron of putative events must overlap an annotated exon, with the predicted 5'end of the last exon exactly matching the 5'end of the overlapping annotated internal exon. If the putative last exon overlaps an annotated first exon, the matching of 5'ends is permitted a 100nt 'slack' to permit last exon capture despite imprecision of the predicted transcript 5'end, which is of secondary importance for our purposes and is a known limitation of transcript assembly tools with short-read RNA-sequencing
- Last exons overlapping annotated last exons must extend the most distal annotated last exon at a locus and exactly match at the 5'end of the annotated exon
- All 'extension' events must extend a known exon by a user-specified minimum distance (default 100nt)

Putative novel last exons are subsequently merged condition-wise into single GTF using a custom script. Filtering and refining of putative last exons for 3'end precision is subsequently performed condition-wise, with the aim of selecting a single representative last exon prediction for a given condition. Firstly, the distance (in either direction) from the 3'end of last exons to the nearest locus reported in the PolyASite 2.0 database[20] is calculated. Any distance below a user-specified distance (default 100nt) is considered a match and retained for downstream analysis. Considering that 3'seq protocols on which the PolyASite database is based can provide nucleotide resolution of poly(A) sites, the 3'ends of matching last exons are updated to the matching site reported in the PolyASite database.

Given that the PolyASite database (as of version 2.0) lacks datasets with TDP-43 depletion and has limited neuronal cell datasets, it is possible it has incomplete coverage of polyadenylation sites specific to TDP-43 depletion and/or neuronal cell contexts. PolyA sites are characterised by an enriched nucleotide distribution around cleavage sites. Most notably, polyA

signal hexamers, of which 18 variants exist[21], are enriched approximately 21 nt upstream of cleavage sites. To rescue 3'ends of last exons that may not be represented in the PolyASite database, the final 100nt of putative last exon sequences are extracted and last exons are retained if an exact match to any of the previously defined 18 polyA signal hexamers[21] is found. If a locus contains multiple predicted last exons, the last exon with a polyA signal hexamer located closest to the expected 21nt upstream distance is selected for each experimental condition.

Next, a combined transcriptome reference of novel and annotated last exons is generated. All last exons passing either the PolyASite or motif filter are retained for downstream analysis. Last exons of all annotated transcripts are extracted using a custom script. Last exons of each gene are assigned a common 'last exon isoform identifier' based on any overlapping sequence. 3'UTR extensions are assigned a distinct identifier, and the annotated last exon(s) they extend are grouped into a single identifier. This has the effect of comparing the usage of 3'UTR extension to all other annotated last exons of the same gene. In order to prevent misattribution of reads to internal last exon isoforms, any regions overlapping annotated first or internal exons are removed (i.e. only 'unique' regions of last exons are retained). Any last exons with a 3'end overlapping annotated first/internal exons are also completely removed from downstream analysis to produce a final reference of last exon isoforms for quantification.

Transcript sequences are extracted using gffread[69] version 0.12.1 and used to produce a decoy aware transcriptome index constructed using Salmon[22] (version 1.5.2) with full genome sequence (Grch38 build) used as decoys[71]. Samples are subsequently quantified against the last exon reference using Salmon[22] v1.5.2 with the '--gcBias' & '--seqBias' flags. Transcript per million (TPM) values of individual transcripts are summed according to their assigned last exon isoform ID, and estimated counts are generated using the 'countsFromAbundance=lengthScaledTPM option' in the tximport[72] package (version 1.26.0). The counts matrix is optionally used as input to DEXSeq[23] v1.44.0 to test for differential isoform usage between experimental conditions. A relative polyA site usage is further calculated for each gene by dividing the expression (in TPM units) of each last exon isoform by the sum of expression of all isoforms of the gene.

PAPA v0.2.0 is implemented as a Snakemake[61] pipeline. All interval operations in Python are performed using the PyRanges[73] package, and genomic sequence operations using a combination of pyfaidx[74] and BioPython[75]. Package dependencies are managed using conda environments, including an 'execution' environment with a minimal Snakemake installation. The code is available at https://github.com/frattalab/PAPA.

Identification of cryptic last exons with PAPA

To generate a common transcript reference against which to quantify last exons, we first ran PAPA in 'identification' mode to predict novel last exons across all stranded datasets in our compendium (All i3Neuron datasets, two of the SH-SY5Y datasets and one SK-N-BE(2) dataset). StringTie[19] transcript assembly was performed in an 'annotation-guided' manner using

a filtered GENCODE v40[58] human transcriptome reference. The reference was filtered first for transcript models with a 'transcript support level' tag value of at least 3 that belong to protein-coding or lncRNA genes. Transcript models with the 'mRNA_end_NF' tag, denoting transcripts with unsupported 3'ends, are also removed as performed by LABRAT[76].

Gene transfer format (GTF) files of predicted last exons across all datasets were subsequently combined into a single GTF of novel last exons using the custom script 'combine_novel_last_exons.py' script available in the PAPA repository. All datasets were then quantified and assessed for differential usage using a unified transcriptome reference of combined novel last exons and annotated last exons (from the same filtered GTF used for transcript assembly). All differential usage tests were performed using the standard DEXSeq workflow without additional covariates, with the exception of the Klim et al i3 motor neuron dataset[47] where the date of differentiation was added as a covariate. Cryptic last exons were defined as isoforms with an Benjamini-Hochberg adjusted p-value of < 0.05 from DEXSeq, mean usage in control samples < 10 % and increase in mean usage in knockdown samples > 10 %.

Following manual inspection of cryptic events we observed frequent IPA calls resulting from regions with intron retention, where the predicted 3'end can be attributed to regions with marked well of reduced coverage (likely due to repetitive sequence) but similar levels of coverage either side of the well and at the intron-exon boundaries. We therefore manually curated cryptic IPA events to mitigate these artefacts. We do not anticipate intronic ALEs to be similarly affected because their 5'end is defined by a novel splice junction.

TDP-43 iCLIP analysis

The SH-SY5Y TDP-43 iCLIP data was generated and processed as previously described[10], and the raw data is available at E-MTAB-11243. iCLIP peaks from the two independent replicates were merged into non-redundant intervals for all subsequent analysis.

Cryptic events are defined as last exon isoforms passing cryptic thresholds in any *in vitro* dataset. The probability of detecting TDP-43 binding events via iCLIP is influenced by the abundance of target RNAs, but by pooling cryptic events across datasets we cannot control for the confounding influence of RNA expression between groups. We therefore defined background events as isoforms that were assessed for differential usage (see above criteria) in all SH-SY5Y datasets and had a padj > 0.05 across all datasets. Given that the cryptic group contains events identified across all datasets (with no guarantees of expression in SH-SY5Y datasets), we therefore penalise the detection of TDP-43 binding in the cryptic group and bias against observing enriched binding in this group.

To define representative intervals for 3'Ext events, the most distal annotated polyA site is selected to represent the proximal site, and background events represent loci with a predicted novel 3'UTR extension. For other event categories, background events include annotated and

novel events. However, our approach to define a common last exon reference across datasets and defining last exon isoforms (see above) can result in non-redundant intervals being predicted for the same last exon isoform. As such, we implemented a collapsing strategy to define a single representative interval for each event. First, overlapping novel predictions are filtered for those that match a site from the PolyASite atlas. If distinct reference sites are reported for the same isoform, the site that is predicted in the most independent datasets is selected as representative. If distinct sites are detected in the same number of independent datasets, the more proximal site is arbitrarily selected as representative. Finally, because the standard PolyASite atlas cluster intervals were used for matching, distinct 3'end predictions can overlap with cluster intervals. In these cases the site that is closest to the PolyASite representative coordinate is selected. If these distinct cluster-overlapping sites are equidistant from the representative coordinate, the most distal coordinate is arbitrarily selected. If no isoforms match an atlas site (i.e. contains a polyA signal sequence), we first attempt to select a representative site whose putative polyA signal motif minimises the deviance from the characteristic position 21nt upstream of the polyA site. If multiple motifs are equidistant from the expected position, the most proximal site is arbitrarily selected as representative. For ALE and IPA events, we noted that collapsing at the 3'end still resulted in distinct intervals for each last exon. 9 background IPA events still had distinct 3'end coordinates following the filtering criteria above, and the interval with the most distal 3'end was arbitrarily selected as the representative interval. 4 background ALE events had distinct 5'end coordinates following the above filtering, and the most proximal 5'end (i.e. shortest exon) is arbitrarily selected as the representative interval. Given the very small proportion of events affected by these arbitrary criteria, we do not anticipate they will substantially affect the analysis.

To construct metaprofiles of TDP-43 binding, Single nucleotide regions representing the genomic landmarks were extended by 500nt in both directions and per-position coverage by iCLIP peaks was then calculated, assigning a value of 1 when a given position overlaps an iCLIP peak. All interval operations were performed using bedtools[77] version 2.31.0. The mean (equivalent to the fraction of events that have an overlapping iCLIP peak) and standard error of coverage is then calculated for each position relative to the landmark. Confidence intervals around the mean coverage in the maps correspond to +/- 1 standard error. Both mean coverage and confidence intervals are visualised following LOESS smoothing with the 'span' parameter set to 0.1.

De-novo motif enrichment analysis

To perform de-novo motif enrichment, we adapted PEKA[24], which identifies kmers with positional enrichment at iCLIP peaks relative to background cross-link sites whilst normalising to the general occurrence in the surrounding genomic context. Therefore, we can substitute iCLIP peaks and global cross-link sites for cryptic and background landmarks respectively to identify positionally enriched kmers with respect to cryptic landmarks. For all comparisons, we ran PEKA to search for enriched 6-mers in the proximal window of interest set to 250nt (the broad window in which iCLIP peaks were observed) and the distal window set to 500nt (to maintain consistency with the overall search space for iCLIP peaks). The 'percentile' flag was set to 0 to

switch off thresholding of background regions based on read count, and the 'relpos' flag to 0 to consider all positions in the proximal window when calculating the enrichment score.

Preferred TDP-43 binding 6-mers were extracted from Halleger et al.[6]. Briefly, The 6-mers were defined using PEKA as the as the top 20 most enriched kmers around intronic iCLIP crosslinks across all WT, A326P, G294A, G335A, M337P, Q331K and a 316del346 GFP-TDP-43 in HEK293 cells. The 20 were subsequently separated into the following three groups based on a gradient of enrichment in WT and G335A TDP-43 with respect to A326 and 316del346 variants and their consensus sequence:

- YG-containing [UG]n 6-mers - UGUGUG, GUGUGU, UGUGCG, UGCGUG, CGUGUG, GUGUGC
- YA-containing [UG]n 6-mers - AUGUGU, GUAUGU, GUGUAU, UGUGUA, UGUAUG, UGCAUG
- AA-containing [UG]n 6-mers - GUGUGA, AAUGAA, GAAUGA, UGAAUG, AUGAAU, GUGAAU, GAAUGU, UUGAAU

Where 'Y' corresponds to a pyrimidine nucleotide. To assess their over-representation among enriched 6-mers relative to cryptic landmarks, we performed a one-sided gene-set enrichment analysis (GSEA) using fgsea[32] version 1.24.0 with default settings for each cryptic landmark. The three 6-mer groups and the union of all three groups were provided as input pathways, and kmers were ranked by their PEKA score. After independent runs for each landmark, Benjamini-Hochberg adjusted p-values were calculated with respect to the all tested landmarks and 6-mer sets and used to evaluate statistical significance.

To generate maps of coverage of specific kmers, we used cv_coverage[78] v1.1.0 (https://github.com/ulelab/cv_coverage) to scan for occurrences of the YG-containing [UG]n 6-mers in a 500nt window around cryptic and background landmarks, disabling weighting the occurrence by cDNA count. For coverage plots, the percentage occurrences of each 6-mer were summed separately for the cryptic and background regions. The percentage occurrences were converted to mean coverages and visualised as described for iCLIP maps.

The adapted PEKA code is available at the 'output_mods' branch of the following forked copy of the PEKA repository https://github.com/SamBryce-Smith/peka. A snakemake pipeline to run PEKA and cv_coverage is available in the 'motifs/peka_snakemake' directory of the 'tdp43-apa' repository.

Post-mortem RNA-seq analysis

FACS-seq data processing

Sequenced reads from FACS-sorted frontal cortex neuronal nuclei[26] and were processed as described in Brown et al[10]. The data are available on the Gene Expression Omnibus at GSE126543.

Quantification of cryptic last exons in post-mortem FACS-seq data

Nuclear RNA-seq libraries contain both nascent and processed RNA. We therefore constructed decoy transcript models that include intron retention at the ALE and IPA loci to limit the confounding effect of nascent RNAs on transcript quantification[22]. Cryptic last exons are first classified as ALE, IPA or 3'Exts using the same criteria as PAPA, and decoy transcript models are subsequently generated separately for each event type.

For IPA events, the unique cryptic last exon sequence is extended to incorporate the annotated internal exon (up to its 5' boundary). Then, a 'spliced' decoy transcript is generated that traverses the annotated internal exon to the downstream exon for all annotated transcripts, and an 'intron retention' decoy transcript is generated that contains the same pairs of internal exons and the intervening intron. For ALEs, a 'retained intron' decoy transcript is generated that consumes the complete intronic sequence in which the last exon is contained. No decoy transcript models are generated for 3'Ext events. Decoy transcript identifiers are appended with suffixes to differentiate from cryptic ALEs and last exon identifiers are generated (excluding decoys) with respect to novel and annotated last exons as in PAPA to allow calculation of % PAU usage. The script used to generate the decoy-augmented last exon reference is available at 'add_decoys_to_gtf.py' in the 'tdp43-apa' GitHub repository.

The decoy-augmented reference quantified with Salmon v1.8.0[22] using the 'salmon' sub-pipeline available at https://github.com/frattalab/rna_seq_single_steps. As with PAPA, samples are quantified against a decoy-aware transcriptome index with full genome sequence (GRCh38 build) used as decoys[71] and the '--gcBias' and '--seqBias' flags enabled.

Calculation of % polyA usage is performed using a copy of the 'tx_to_polyA_quant.R' script from the PAPA repository. Sample-wise differences in % polyadenylation site usage is calculated by subtracting the usage in the TDP-43 positive population from the TDP-43 negative population, such that a positive difference indicates the cryptic APA has a higher relative expression in the population with TDP-43 depletion.

New York Genome Centre (NYGC) RNA-seq data

The sequencing libraries were generated[27,79] and processed[13] as previously described. Samples were classified into disease subtypes as previously described[13]. Briefly, FTD subtypes were classified by pathology according to the presence of TDP-43 inclusions (FTLD-TDP), FUS or Tau aggregates. ALS patients were sub-categorised based on presence (ALS-non-TDP) or absence (ALS-TDP) of reported SOD1 or FUS mutations. The following samples were considered as regions where TDP-43 pathology (and specific cryptic junction expression) is expected; motor (ALS-TDP), frontal and temporal cortex samples (FTLD-TDP, ALS-TDP), cervical, lumbar and thoracic spinal cord samples (ALS-TDP).

We opted to quantify ALE events using junction reads, which provide direct quantification of the occurrence of a splicing event. As of version 0.2 PAPA does not directly report splice junctions associated with ALE events. However, as the filtering criteria applied by PAPA requires putative ALE events to have a terminal splice junction with a direct match to an annotated 5'ss, it is

possible to infer splice junctions from reference annotation using just the reported last exon coordinates. For ALEs fully contained within annotated introns, the splice junction is defined from the intron start to the start of the ALE. If last exons are distal to the annotated gene, then the closest upstream annotated intron is found. The splice junction is subsequently defined as the region from intron start to the start of the ALE. Finally, for annotated ALEs all annotated introns that terminate at the ALE are reported as splice junctions for the event. The above steps are implemented in a custom script 'last_exons_to_sj.py' available at the 'tdp43-apa' GitHub repository.

Splice junctions for cryptic ALEs and cryptic splice junctions identified in cortical-like i3Neurons[13] were quantified across the NYGC RNA-seq cohort by extracting counts for provided junctions from the '.SJ.out.tab' files produced by STAR[59]. The code is implemented in the 'bedops_parse_star_junctions' v0.1.0 Snakemake pipeline and is available at https://github.com/SamBryce-Smith/bedops_parse_star_junctions.

We defined detection criteria to prioritise cryptic splice junctions that are specifically in tissue types and samples with expected TDP-43 pathology. Junctions are considered expressed if at least two spliced reads are detected in a sample. Junctions are considered selectively expressed if expressed in at most 0.5 % of all samples where TDP-43 pathology is not expected and at least 1 % of samples where TDP-43 pathology is expected. We note that such criteria will exclude events with enriched expression in tissues with expected TDP-43 proteinopathy, but that have basal expression in unknown cell types not represented in our *in vitro* compendium. Such events may still have relevance in mechanisms of disease in specific cell types, but are less suitable for discriminating samples with TDP-43 proteinopathy.

Ribo-seq analysis

i3Neuron Ribo-seq data was generated and processed as previously described[13]. Uniquely mapped reads were assigned to genes based on the union of annotated 'CDS' entries in the Gencode v34 standard annotation released using featureCounts[80] version 2.0.1. Differential expression between TDP-43 knockdown and control was performed using DESeq2[81] v1.38.3, and differentially translated genes were defined based on a Benjamini-Hochberg adjusted p-value threshold of 0.05. Any last exon passing our cryptic criteria in at least one of the i3 Neuron datasets (Brown i3 cortical, Seddighi i3 cortical, Humphrey i3 cortical) was considered for intersection with differentially translated genes.

Gene-set enrichment analysis was performed using fgsea[32] version 1.24.0 with default settings. Cryptic 3'Ext, IPA & ALE containing genes were provided as input pathways, and moderated fold changes calculated with 'lfcShrink' function from DESeq2 package using the default apeglm[82] method as the shrinkage estimator to rank genes. A threshold of 0.05 Benjamini-Hochberg adjusted p-value was used to determine statistical significance.

Read counting was performed using the 'feature_counts' sub-pipeline available at https://github.com/frattalab/rna_seq_single_steps. Custom scripts used to perform differential expression and pathway analysis are available at https://github.com/frattalab/tdp43-apa .

For cross-referencing with differential RNA expression, we used differential expression analysis from cortical-like i3Neurons performed as previously described[13]. Cryptic last exon-containing genes were highlighted if they passed the statistical significance threshold in the Ribo-seq differential expression analysis.

Analysis of ELK1 transcription factor activity

ELK1 target genes in HeLa cells were accessed from the ChIP-Atlas[83] on the 15th November 2023. We used the 'Target genes' module to obtain a list of target genes that have a ChIP-seq peak within +/- 1 kilobase of transcription start sites. The resulting list contained two HeLa datasets (GSM608163, GSM935326), and was filtered to target genes identified in both datasets Given a reported redundancy of function between ELK1 and other members of the ternary complex factor family[84] (ELK3 and particularly ELK4), we also attempted to define a unique set of ELK1 target genes. ELK4 target genes in HeLA cells were accessed from ChIP-Atlas on the 29th November 2023 using the same parameters. The resulting list contained 3 HeLa datasets (GSM608161, GSM608162, GSM935351), and we again filtered for target genes identified in all datasets. ELK3 HeLa ChIP-seq data was not available through ChIP-Atlas at the time of publication, and was not considered for further redundancy. ELK3 RNA levels are 10x lower than ELK3 and ELK4 in HeLa TDP-43 knockout cells[39], so we anticipate this is unlikely to affect our conclusions. ELK1 and ELK4 target gene lists were intersected to define common and unique target genes for each transcription factor. Final target gene lists used are reported in Supplementary Table 5.

RNA-sequencing data from HeLa cells with TDP-43 knockout[39] were accessed from GSE136366. The data were processed and differential expression was performed as described in Brown et al[10]. Genes were ranked by DESeq2's test statistic ($\log_2$ transformed fold change divided by the standard error of the fold change) after removing genes with differential splicing upon TDP-43 knockout, where we can expect to attribute any changes in gene expression to TDP-43 loss of function. Differentially spliced genes were defined using MAJIQ[85], considering any genes with a probability of > 0.95 as differentially spliced. The target gene sets described above were used as input pathways to fgsea[32] version 1.24.0 using default settings.

Code availability

All visualisation and statistical testing was performed in R[86] version 4.3.2 using the ggplot2[87] v3.4.4, ggpubr[88] v0.6.0, ggprism[89] v1.0.4 and ggrepel[90] v0.9.4 packages. Preprocessing for visualisation and generation of supplementary tables was performed using tidyverse[91] v2.0.0, writexl[92] 1.4.2, and data.table[93] v1.14. Unless otherwise stated, analyses requiring genomic interval operations or queries with bioinformatics data formats were performed in Python 3.10.11 using PyRanges[73] 0.0.127, pandas[94] v2.0.2 numpy[95] v1.23. Analysis and visualisation code,

along with conda[96] and renv[97] environments for dependency management, can be accessed at https://github.com/frattalab/tdp43-apa. Alternative repositories for specific analyses are reported in the relevant sections of the methods.

Author contributions

Conceptualization: PF, MS, SB-S
Data Curation: SB-S, ALB, AM, SB, MZ, OGW, YW, JH
Formal analysis: SB-S, A-LB, PRM, FM, AM, SB
Funding acquisition: PF,MS, MEW, TR
Investigation: SB-S,A-LB,PRM,FM,AM,MY,SB,YA.Q,SH,JNV,KS,ER, MEW
Methodology: SB-S,A-LB,PRM,FM,SE.H,MZ,MK,OGW,JH,NB,MS,PF
Project administration: PF, MS
Resources: JH,MW,PF, TR
Software: SB-S,A-LB,AM
Supervision: PF,MS
Visualisation: SB-S, A-LB, PRM, AM, SB
Writing - Original Draft: SB-S,PF,MS
Writing - Review & Editing: All authors

Acknowledgements

SB-S is supported by a UK Motor Neurone Disease Association and Masonic Charitable Foundation PhD Studentship (893792). PF 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 U54NS123743, Target ALS and The Robert Packard Center for ALS Research. MS is supported by a UKRI Future Leaders Fellowship (MR/T042184/1). PRM is supported by a Wellcome Trust Clinical Training Fellowship (102186/B/13/Z). This work was supported, in part, by the Intramural Research Program of the National Institutes of Neurological Disorders and Stroke (MEW), by the Center for Alzheimer's and Related Dementias, National Institute on Aging and National Institute of Neurological Disorders and Stroke (MEW), by the Robert Packard Center for ALS Research (MEW).

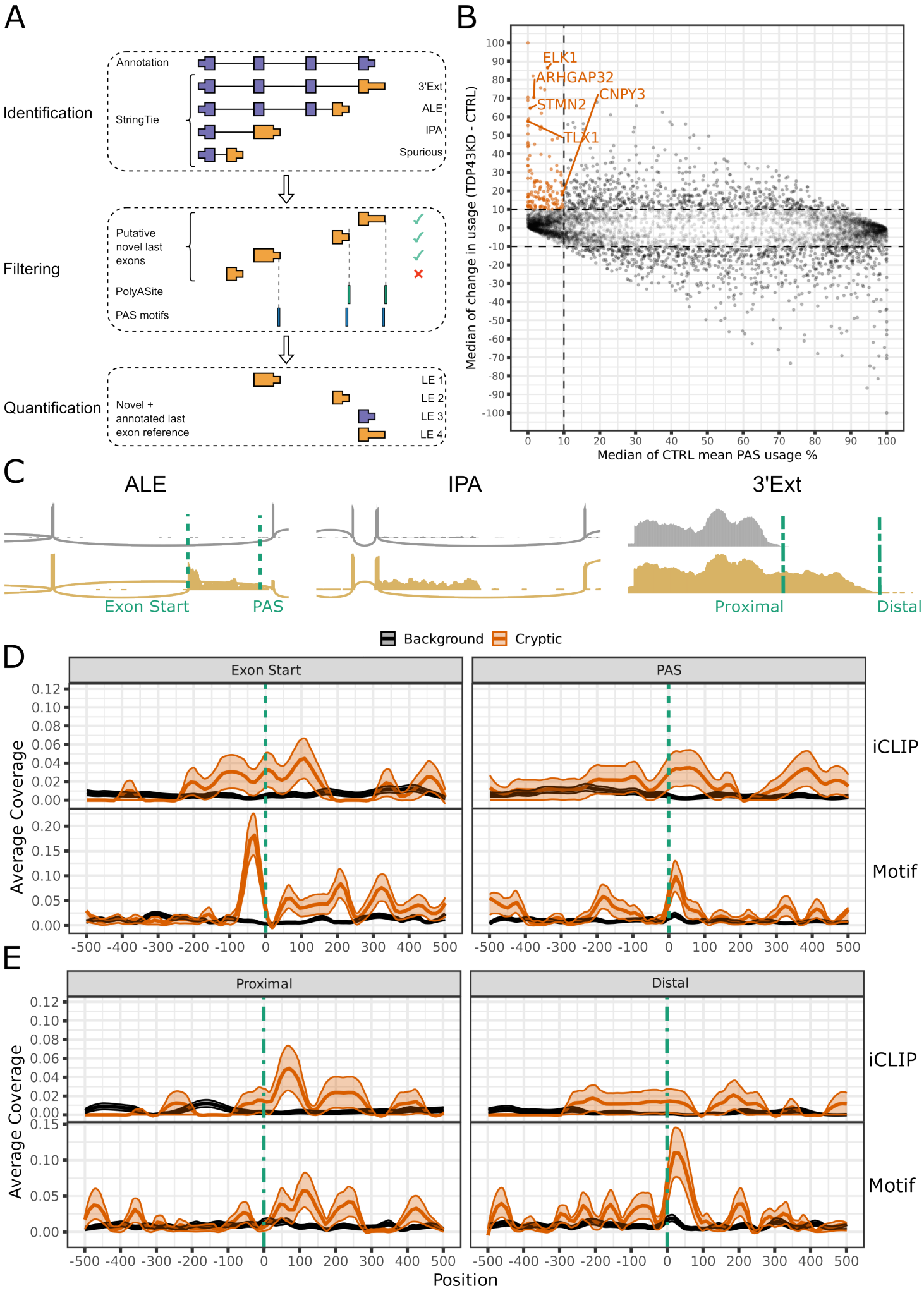

A
Identification
Annotation
StringTie
3'Ext
ALE
IPA
Spurious
Filtering
Putative novel last exons
PolyASite
PAS motifs
Quantification
Novel + annotated last exon reference
LE 1
LE 2
LE 3
LE 4
B
ELK1
ARHGAP32
CNPY3
STMN2
TLX1
Median of change in usage (TDP43KD - CTRL)
Median of CTRL mean PAS usage %
C
ALE
IPA
3'Ext
Exon Start
PAS
Proximal
Distal
Background
Cryptic
D
Exon Start
PAS
Average Coverage
iCLIP
Motif
E
Proximal
Distal
Position

Figure 1

Figure 1 - TDP-43 depletion induces cryptic APA in a compendium of in vitro TDP-43 datasets

A) Schematic demonstration of the computational pipeline to detect, quantify and infer differential usage of last exons from bulk RNA-seq data. De-novo transcripts are assembled using StringTie[19] and subsequently filtered for a mean TPM > 1 within each experimental condition. Last exons are extracted from transcript models and compared to reference annotation (purple) to identify putative novel last exons (orange). Putative last exons are filtered for predicted 3'ends < 100nt from sites in the PolyASite[20] database, or rescued if a conserved polyA signal hexamer[21] can be identified within the last 100nt of the last exon. Novel and annotated last exons were subsequently quantified using Salmon[22] and assessed for differential usage using DEXSeq[23]. For further details see Methods.

B) Last exons responsive to TDP-43 depletion. All points represent a last exon passing a Benjamini-Hochberg adjusted p-value < 0.05 threshold in at least one dataset. Where a last exon passes the threshold in multiple datasets, the median values across datasets are calculated to represent the basal usage and change in usage upon TDP-43 depletion. Last exons passing our cryptic threshold are highlighted in orange (Benjamini-Hochberg adjusted p-value < 0.05, mean usage in control cells < 10 % and change in usage between TDP-43 knockdown and control ('TDP43KD' - 'CTRL') > 10 %).

C) Example RNA-seq coverage traces control (grey) and TDP-43 knockdown (gold) i3Neuron samples for cryptic ALE (*ARHGAP32*), IPA (*ANKRD27*) and 3'Ext (*TLX1*) events. Dashed lines indicate landmarks around which TDP-43 binding is assessed in (C) and (D). *ARHGAP32* and *ANKRD27* are encoded on the reverse strand but flipped to read 5'-3' for visualisation purposes.

D) TDP-43 binding maps around boundaries of ALEs.
(Top) TDP-43 iCLIP RNA maps around the first nucleotide of the last exon ('Exon Start') and the polyA site ('PAS') of ALEs. The solid lines represent the mean coverage (equivalent to fraction coverage) of relative positions upstream (negative values) and downstream (positive values) of the landmark from pooled TDP-43 iCLIP peaks from SH-SY5Y cells (n=2) in background (black, n=929) and cryptic ALEs (orange, n=92). Shaded intervals represent +/- 1 SE for the corresponding coloured group.
(Bottom) TDP-43 motif maps around the first nucleotide of the last exon ('Exon Start') and the polyA site ('PAS') of ALEs. The solid lines represent the mean coverage by YG-containing hexamers (Supplementary Fig. 3A) of relative positions upstream (negative values) and downstream (positive values) of the landmark in background (black, n=929) and cryptic ALEs (orange, n=92).

E) TDP-43 binding maps around alternative polyA sites of 3'Exts.
(Top) TDP-43 iCLIP RNA maps around the proximal ('Proximal') and distal ('Distal') polyA site of 3'Ext events. As in D), but background (black, n=798) and cryptic regions (orange, n=86) are obtained for 3'Ext events.
(Bottom) TDP-43 motif maps around the proximal ('Proximal') and distal ('Distal') polyA site of 3'Ext events. As in D), but background (black, n=798) and cryptic regions (orange, n=86) are obtained for 3'Ext events.

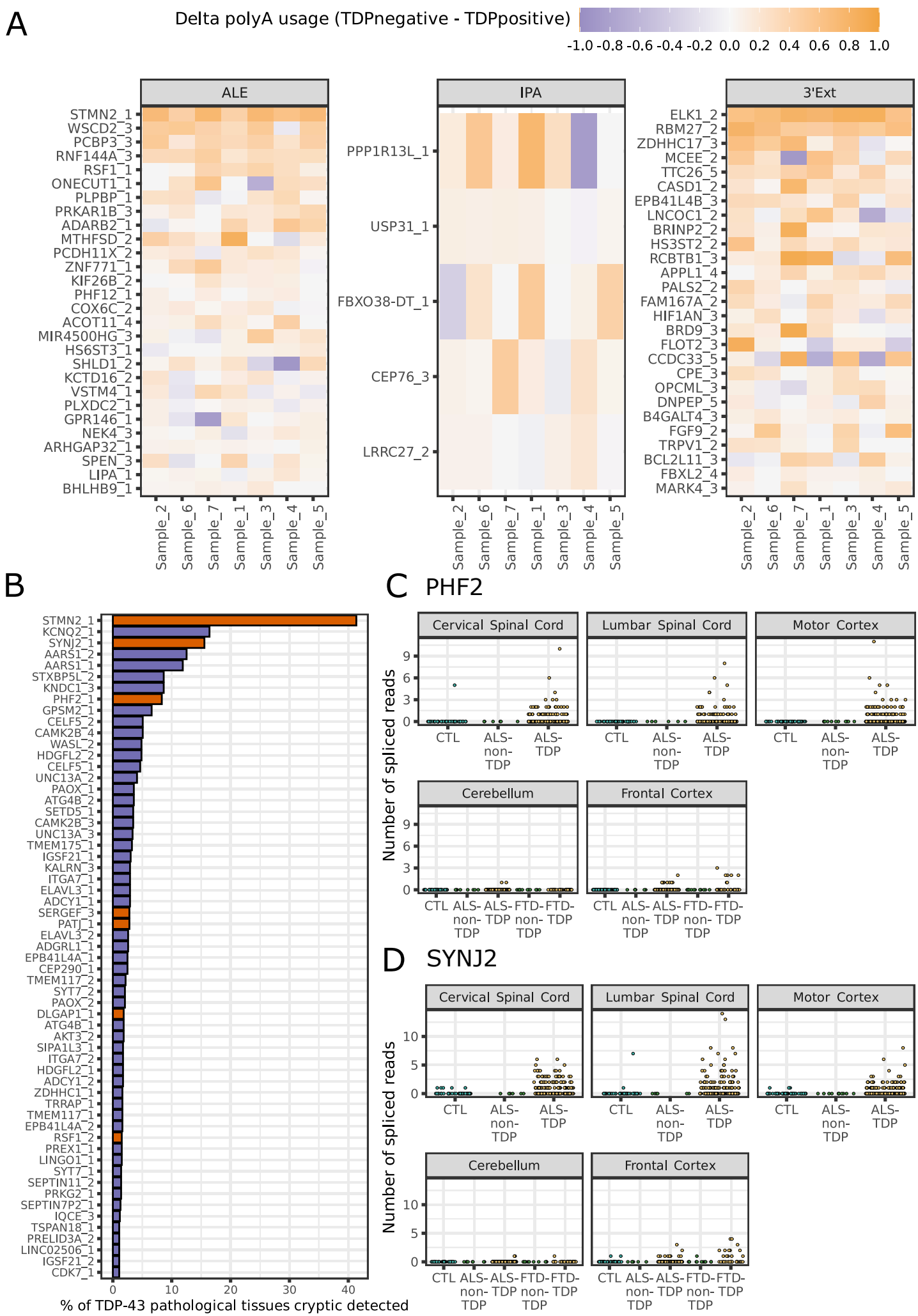

A
Delta polyA usage (TDPnegative - TDPpositive)
-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
ALE
IPA
3'Ext
STMN2_1
WSCD2_3
PCBP3_3
RNF144A_3
RSF1_1
ONECUT1_1
PLPBP_1
PRKAR1B_3
ADARB2_1
MTHFSD_2
PCDH11X_2
ZNF771_1
KIF26B_2
PHF12_1
COX6C_2
ACOT11_4
MIR4500HG_3
HS6ST3_1
SHLD1_2
KCTD16_2
VSTM4_1
PLXDC2_1
GPR146_1
NEK4_3
ARHGAP32_1
SPEN_3
LIPA_1
BHLHB9_1
PPP1R13L_1
USP31_1
FBXO38-DT_1
CEP76_3
LRRC27_2
ELK1_2
RBM27_2
ZDHHC17_3
MCEE_2
TTC26_5
CASD1_2
EPB41L4B_3
LNCOC1_2
BRINP2_2
HS3ST2_2
RCBTB1_3
APPL1_4
PALS2_2
FAM167A_2
HIF1AN_3
BRD9_3
FLOT2_3
CCDC33_5
CPE_3
OPCML_3
DNPEP_5
B4GALT4_3
FGF9_2
TRPV1_2
BCL2L11_3
FBXL2_4
MARK4_3
Sample_2
Sample_6
Sample_7
Sample_1
Sample_3
Sample_4
Sample_5
B
STMN2_1
KCNQ2_1
SYNJ2_1
AARS1_2
AARS1_1
STXBP5L_2
KNDC1_3
PHF2_1
GPSM2_1
CELF5_2
CAMK2B_4
WASL_2
HDGFL2_2
CELF5_1
UNC13A_2
PAOX_1
ATG4B_2
SETD5_1
CAMK2B_3
UNC13A_3
TMEM175_1
IGSF21_1
KALRN_3
ITGA7_1
ELAVL3_1
ADCY1_1
SERGEF_3
PATJ_1
ELAVL3_2
ADGRL1_1
EPB41L4A_1
CEP290_1
TMEM117_2
SYT7_2
PAOX_2
DLGAP1_1
ATG4B_1
AKT3_2
SIPA1L3_1
ITGA7_2
HDGFL2_1
ADCY1_2
ZDHHC1_1
TRRAP_1
TMEM117_1
EPB41L4A_2
RSF1_2
PREX1_1
LINGO1_1
SYT7_1
SEPTIN11_2
PRKG2_1
SEPTIN7P2_1
IQCE_3
TSPAN18_1
PRELID3A_2
LINC02506_1
IGSF21_2
CDK7_1
0
10
20
30
40
% of TDP-43 pathological tissues cryptic detected
C PHF2
Cervical Spinal Cord
Lumbar Spinal Cord
Motor Cortex
Cerebellum
Frontal Cortex
Number of spliced reads
0
3
6
9
CTL
ALS-non-TDP
ALS-TDP
FTD-non-TDP
FTD-TDP
D SYNJ2
Cervical Spinal Cord
Lumbar Spinal Cord
Motor Cortex
Cerebellum
Frontal Cortex
Number of spliced reads
0
5
10
CTL
ALS-non-TDP
ALS-TDP
FTD-non-TDP
FTD-TDP

Figure 2

Figure 2 - Cryptic last exons are detected in post-mortem ALS-FTD RNA-seq datasets

A) Heatmap of cryptic polyadenylation site usage in post-mortem FACS-seq data[26]. Cells are coloured according to the magnitude of sample-wise difference in usage between TDP-43 depleted (TDPnegative) and TDP-43 positive (TDPpositive) cells. Rows represent individual cryptic last exons from in-vitro that passed enrichment criteria (median sample-wise difference in usage (TDPnegative - TDPpositive) > 5 %) and are arranged in descending order of the difference in usage within each event type. Columns represent individual patients within the cohort.

B) Selectively expressed cryptic ALEs (orange) and splicing events[13] (purple) in tissues and samples with TDP-43 proteinopathy in the New York Genome Centre (NYGC) ALS Consortium dataset. Events are considered detected if at least 2 junction reads were detected in a sample.

C) Detection of spliced reads for the cryptic ALE in *PHF2* across samples in the NYGC ALS Consortium dataset. 'CTL' denotes control samples. Colour indicates whether disease subtype and region is expected (orange) or not expected (green) to have TDP-43 pathology and cryptic spliced read expression.

D) As in C), but for cryptic ALE in *SYNJ2*.

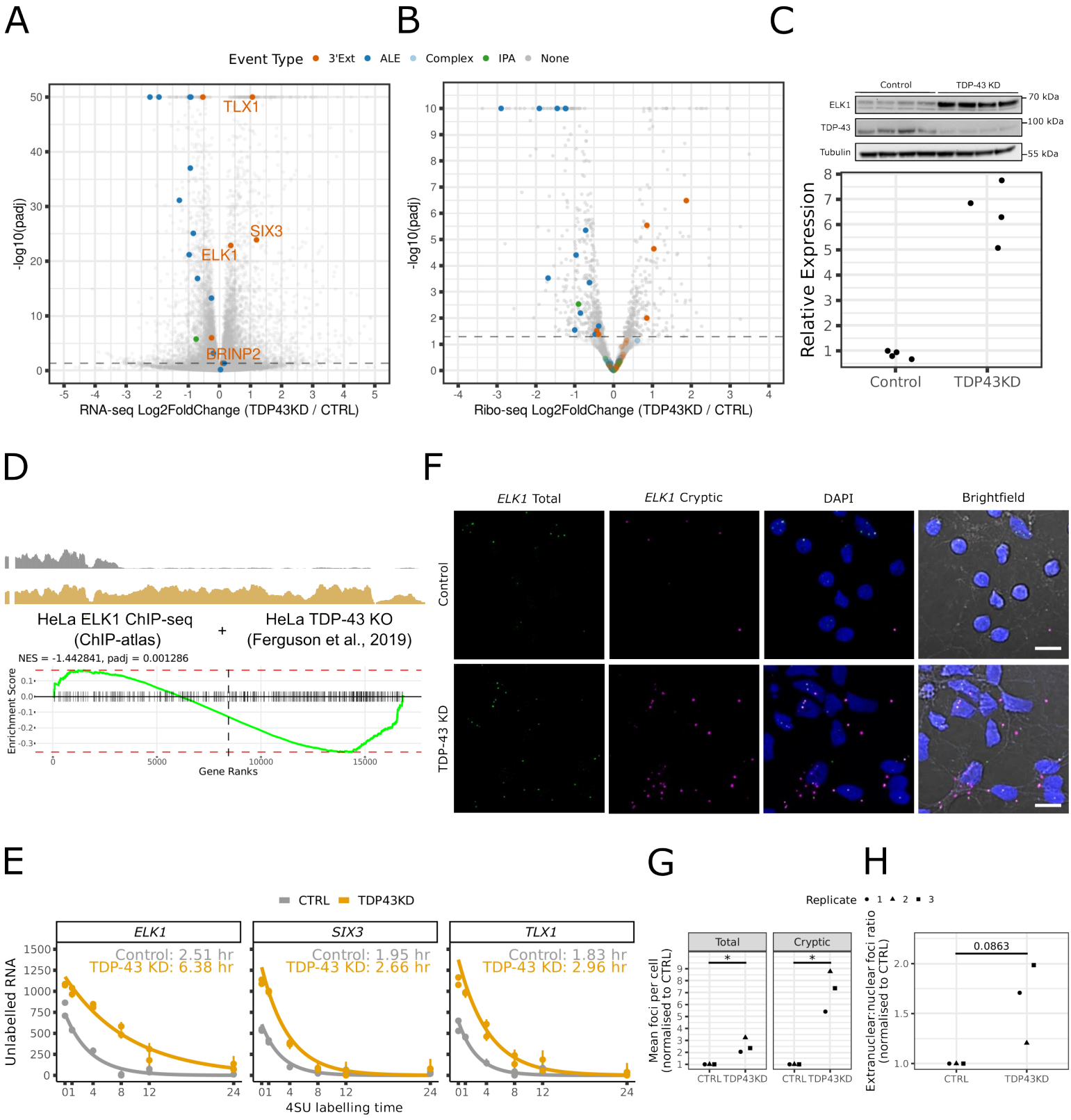

A
B
C
Event Type 3'Ext ALE Complex IPA None
TLX1
SIX3
ELK1
BRINP2
-log10(padj)
RNA-seq Log2FoldChange (TDP43KD / CTRL)
Ribo-seq Log2FoldChange (TDP43KD / CTRL)
Control
TDP-43 KD
ELK1
TDP-43
Tubulin
70 kDa
100 kDa
55 kDa
Relative Expression
Control
TDP43KD
D
HeLa ELK1 ChIP-seq (ChIP-atlas)
+
HeLa TDP-43 KO (Ferguson et al., 2019)
NES = -1.442841, padj = 0.001286
Enrichment Score
Gene Ranks
F
ELK1 Total
ELK1 Cryptic
DAPI
Brightfield
Control
TDP-43 KD
E
CTRL TDP43KD
ELK1
SIX3
TLX1
Control: 2.51 hr
TDP-43 KD: 6.38 hr
Control: 1.95 hr
TDP-43 KD: 2.66 hr
Control: 1.83 hr
TDP-43 KD: 2.96 hr
Unlabelled RNA
4SU labelling time
G
Total
Cryptic
Mean foci per cell (normalised to CTRL)
CTRL TDP43KD
H
Replicate 1 2 3
0.0863
Extranuclear:nuclear foci ratio (normalised to CTRL)
CTRL
TDP43KD

Figure 3

Figure 3 - Cryptic 3'UTR extensions in transcription factor RNAs lead to increased RNA and protein levels by increased RNA stability and altered localisation

A) Volcano plot of differential expression analysis of RNA-seq data between TDP-43 knockdown (TDP43KD) and control (CTRL) i3Neurons. Cryptic 3'Ext genes with increased translation (Fig. 3B) are highlighted in orange. Genes with a $-\log_{10}$ transformed Benjamini-Hochberg adjusted p-value greater than 50 are collapsed to 50 for visualisation purposes.

B) Volcano plot of differential expression analysis of Ribo-seq data between TDP-43 knockdown (TDP43KD) and control (CTRL) i3Neurons. Genes are highlighted if they contain a cryptic 3'Ext (orange), ALE (blue) or IPA (green) event. Genes with a $-\log_{10}$ transformed Benjamini-Hochberg adjusted p-value are collapsed to 10 for visualisation purposes.

C) Western blot analysis of ELK1 protein levels in Halo-TDP-43 i3Neurons.
(Top) Western blot showing increased ELK1 protein expression upon TDP-43 KD in Halo-TDP-43 i3Neurons.
(Bottom) Quantification of ELK1 band intensities normalised to Tubulin in control (CTRL) and TDP-43 knockdown (TDP43KD) Halo-TDP-43 i3Neurons.

D) Analysis of ELK1 transcription factor activity.
(Top) Coverage trace in control (black) and TDP-43 knockout (gold) samples for the *ELK1* cryptic 3'Ext in HeLa cells[39].
(Bottom) Enrichment plot for ChIP-seq defined ELK1 target genes in TDP-43 knockout HeLa cells. The green line corresponds to GSEA's running sum statistic, and red horizontal dashed lines mark the maximal enrichment score among upregulated and downregulated genes. The vertical dashed line demarcates the rank between upregulated and downregulated genes in the evaluated gene set. Black vertical bars correspond to locations of ELK1 target genes in the ranked gene set. The black text reports the normalised enrichment score ('NES', normalised to the mean enrichment score of random samples of the same size as the gene set) and Benjamini-Hochberg adjusted p-value ('padj')

E) Decay curve for RNA produced before 4SU labelling in i3Neuron SLAM-seq data for control (grey) and knockdown (orange) samples. Solid curves indicate the fitted estimate of the level of old RNA for each condition. Individual samples are shown as points with the upper and lower 95% credible interval shown as error bars. GrandR-estimated half-lives for control (grey) and knockdown (orange) samples are reported in the inset text for each gene.

F) Representative images for FISH probes targeting the annotated ('*ELK1* Total', green) 3'UTR and cryptic 3'UTR specific ('*ELK1* Cryptic', magenta) sequences of *ELK1* in control (top row) and TDP-43 knockdown (bottom row) i3Neurons. The white lines in the 'Brightfield' column represent scale bars (10 μm).

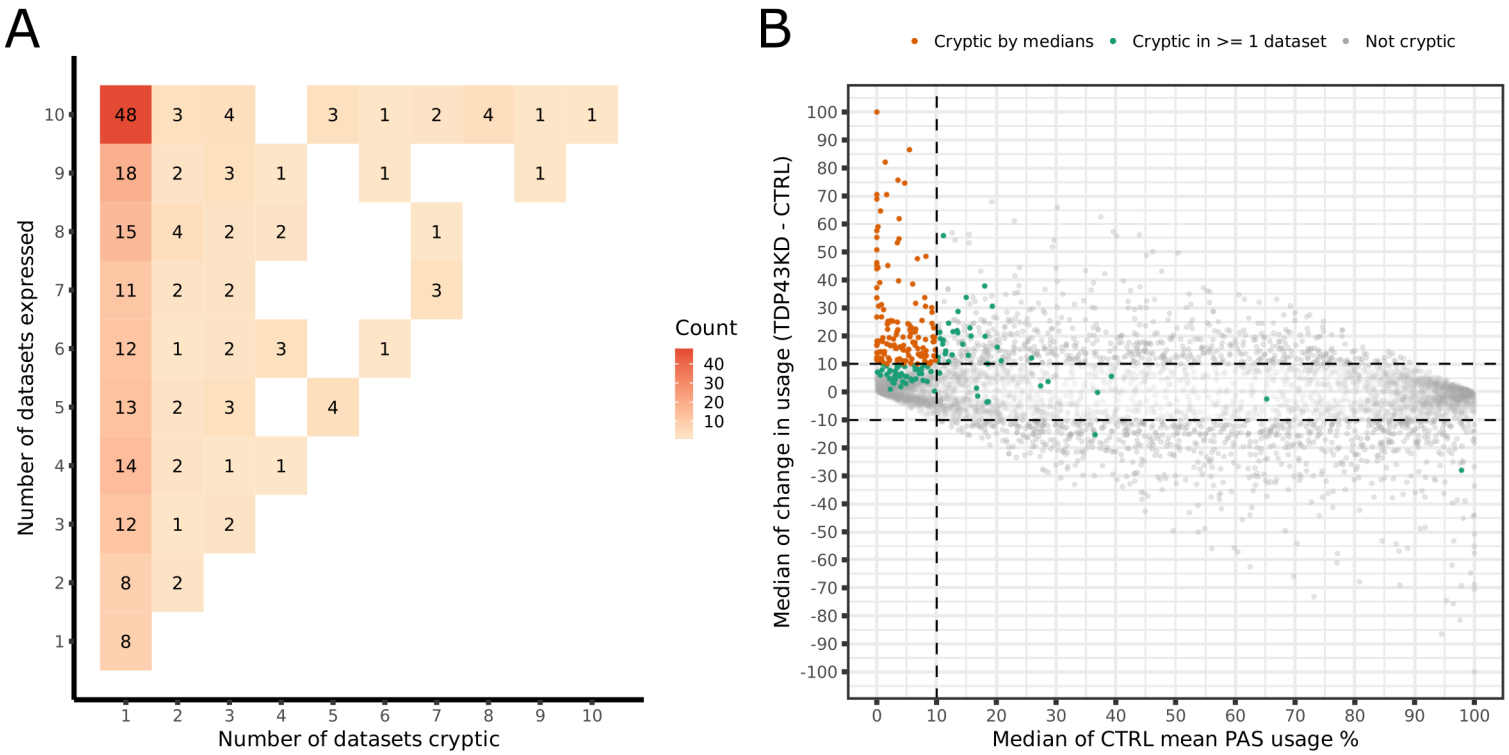
G) Quantification of total FISH probe signal for the probes targeting the annotated 3'UTR ('Total') and cryptic 3'UTR specific ('Cryptic') sequences of *ELK1*. Different shapes represent independent replicates. Mean foci counts per cell (n=10 images) are normalised with respect to the control sample from the same replicate (Methods). A single asterisk (*) represents a Benjamini-Hochberg adjusted p-value < 0.05 from a one-sample t-test on log-transformed ratios (Methods).

H) Subcellular quantification of FISH signal for probes targeting the annotated 3'UTR region ('Total *ELK1*') of *ELK1*. Different shapes represent independent replicates. The mean ratio of extranuclear:nuclei foci counts (n=10 images) is normalised with respect to the control sample from the same experimental replicate (Methods). The numeric label represents the Benjamini-Hochberg adjusted p-value from a one-sample t-test on log-transformed ratios (Methods).

Supplementary Figure 1

Supplementary Figure 1 - Consistency of response to TDP-43 depletion across compendium of in-vitro datasets

Differential usage of cryptic APA events across the compendium of in-vitro datasets. Cells are coloured in accordance to magnitude and direction of change in usage, where positive values (orange) indicate increased usage in TDP-43 knockdown ('TDP43KD') samples. Blank cells indicate the event was not expressed at sufficient levels to be assessed for differential usage. Rows are sorted in decreasing order of the sum of $-\log_{10}$ transformed p-values weighted by the change in usage between TDP-43 knockdown and control samples (TDP43KD - CTRL) in each dataset. A single asterisk indicates that the isoform was considered significantly regulated in a dataset (Benjamini-Hochberg adjusted p-value < 0.05), and two asterisks indicate the isoform is considered cryptic in a given dataset.



Supplementary Figure 2

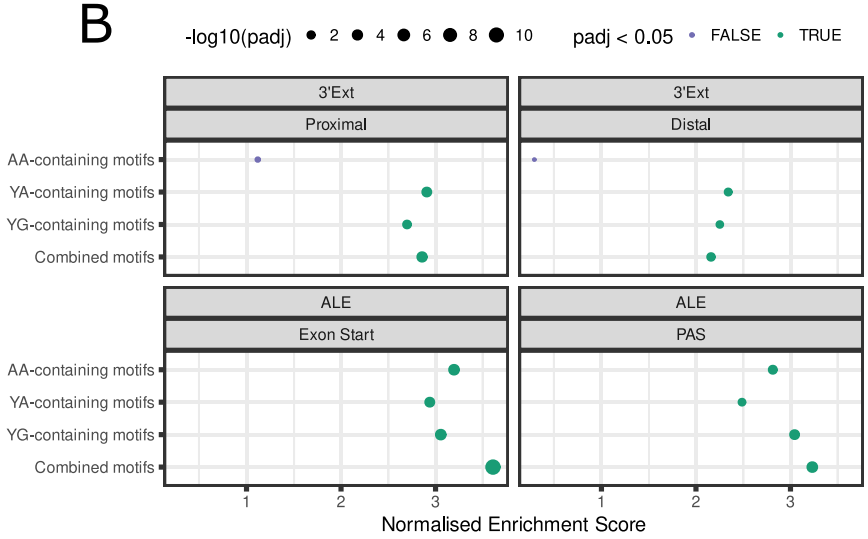
Supplementary Figure 2 - Consistency of cryptic status across compendium of *in vitro* datasets

A) Relationship between the number of datasets in which APAs are called cryptic and their detection. Count labels indicate the number of unique cryptic APAs that fall into the given bin. Events are considered expressed if they pass minimum expression criteria to be evaluated for differential isoform usage (Methods).
B) Last exons responsive to TDP-43 depletion. All points represent a last exon passing a Benjamini-Hochberg adjusted p-value < 0.05 threshold in at least one dataset. Where a last exon passes the threshold in multiple datasets, the median values across datasets are calculated to represent the basal usage and change in usage upon TDP-43 depletion. Points that pass cryptic expression criteria in at least one dataset but pass (orange) or fail (green) the criteria when calculating the median change in usage and expression in control (CTRL) cells across datasets with an Benjamini-Hochberg adjusted p-value < 0.05 are highlighted.

A

AA-containing motifs	GUGUGA, AAUGAA, GAAUGA, UGAAUG, AUGAAU, GUGAAU, GAAUGU, UUGAAU
YA-containing motifs	AUGUGU, GUAUGU, GUGUAU, UGUGUA, UGUAUG, UGCAUG
YG-containing motifs	UGUGUG, GUGUGU, UGUGCG, UGCGUG, CGUGUG, GUGUGC
Combined motifs	All of above

B

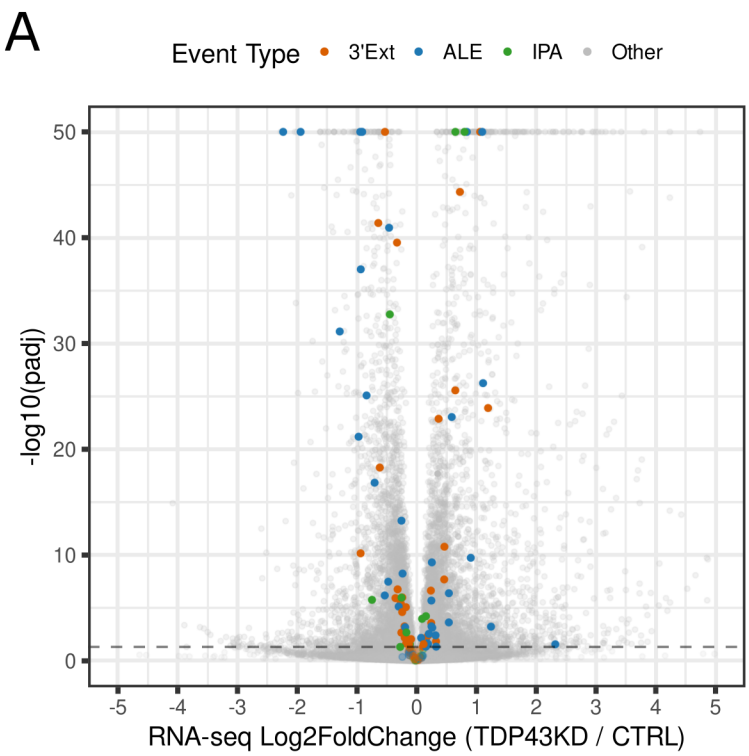


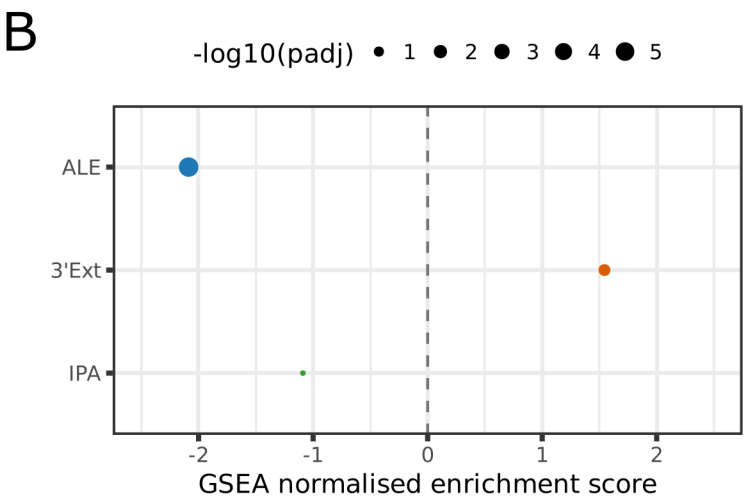
Supplementary Figure 3

Supplementary Figure 3 - Enrichment of previously defined TDP-43 binding hexamers at cryptic APA boundaries

A) Table listing previously defined TDP-43 hexamer groups[6]. 'Y' codes for a pyrimidine nucleotide.

B) Gene set enrichment analysis (GSEA) of enriched TDP-43 binding 6mers on de-novo enriched 6-mers around cryptic landmarks. The panels and labels correspond to regions evaluated for iCLIP binding as in Fig. 1D . The area of the points is proportional to the $-\log_{10}$ transformed adjusted p-value (adjusted with respect to all region types and motif groups), and the colour denotes whether the Benjamini-Hochberg adjusted p-value passes (green) or fails (purple) a significance threshold of < 0.05.





Supplementary Figure 4

Supplementary Figure 4 - Analysis of cryptic APA categories in Ribo-seq data

A) Volcano plot of differential expression analysis of RNA-seq data between TDP-43 knockdown (TDP43KD) and control (CTRL) i3Neurons. Cryptic APA genes with significant differential expression (Benjamini-Hochberg adjusted p-value < 0.05) are highlighted in orange (3'Ext), blue (ALE) or green (IPA). Genes with a $-\log_{10}$ transformed Benjamini-Hochberg adjusted p-value greater than 50 are collapsed to 50 for visualisation purposes.
B) Gene Set Enrichment Analysis (GSEA) of cryptic APA categories in i3Neuron Ribo-seq differential expression fold change ranks. The area of the points is proportional to the $-\log_{10}$ transformed Benjamini-Hochberg adjusted p-value. Points are coloured according to their APA category as in A).