# TDP43 proteinopathy exhibits disease, tissue, and context-specific cryptic splicing signatures


## Abstract

Mislocalization of the nuclear protein TAR DNA-binding protein 43 (TDP43) is a hallmark of ALS and FTD which leads to de-repression and inclusion of cryptic exons. These cryptic exons represent promising biomarkers of TDP43 pathology in a spectrum of neurodegenerative diseases. However, most cryptic exons to date have been identified from *in vitro* models or a single cortical FTD dataset, limiting our understanding of tissue and cell-specific splices as well as differences between *in vitro* and *in vivo* processes. We meta-analyzed published bulk RNAseq datasets representing 1,778 RNAseq profiles of ALS and FTD post-mortem tissue, and *in vitro* models with experimentally depleted TDP43. We identified 142 cryptic splices, including 76 novel events, and identified cryptic splicing signatures with distinct cortical and spinal cord enrichment, among other context-specific profiles. Using RNAseq and RT-qPCR, we validated a subset of these splices in an independent spinal cord cohort, and demonstrated a correlation of TDP43 pathology severity with degree of cryptic splicing. Leveraging multiple public single-nucleus RNAseq datasets of ALS and FTD motor and frontal cortex, we confirmed the elevation of cortical-enriched splices in disease and localized them to layer-specific neuronal populations. This catalog of cryptic splices could inform efforts to develop biomarkers for tissue-specific and cell type-specific TDP43 pathology.


## Introduction

Abnormal processing of the protein TDP43 is observed in post-mortem tissue from a spectrum of neurodegenerative diseases, including amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)[16]. Although normally localized to nuclei, in pathological contexts this splicing factor accumulates in primarily cytoplasmic aggregates. To understand the molecular consequences of nuclear TDP43 loss-of-function, several groups have performed RNAseq profiling on *in vitro* models with experimentally depleted TDP43 levels [5, 14, 17, 23]. These studies identified the accumulation of mRNA transcripts including so-called “cryptic” splice junctions not normally included in mature

transcripts [17, 22]. Some of these cryptic splices, namely in the genes STMN2 and UNC13A, were subsequently observed in RNAseq profiling of FTD and ALS post-mortem tissue, but not healthy controls [5, 14, 20, 23], and were enriched in TDP43-deficient neuronal nuclei sorted from ALS/FTD frontal cortex [18]. This represents powerful evidence demonstrating that the loss of TDP43 splicing repression results in the accumulation of cryptic splices and is a defining pathological hallmark of these proteinopathies.

However, it is not yet possible to directly diagnose TDP43 pathology in nervous system tissues from accessible patient biosamples. Thus, regardless of whether cryptic-spliced mRNAs contribute to the pathogenesis of ALS or FTD, or are merely consequences of a primary proteinopathy, they are potentially of high value as biomarkers. Given their virtual absence in healthy conditions, the detection of even low levels of cryptic mRNA or translated protein in biofluids could facilitate diagnosis and disease monitoring in patients. Furthermore, such biomarkers could enable accurate screening for patient inclusion (or exclusion), or pharmacodynamic evaluation for therapeutics targeting TDP43 pathology. Pharmacodynamic biomarkers could be particularly impactful to enable validation of TDP43-modifying agents in smaller cohorts preceding large, expensive trials

An ideal biomarker assay would likely measure, via RNA or encoded peptide, the levels of one or a small number of all of the possible cryptic splices. Indeed, recent studies have shown the presence of translated cryptic peptides in ALS cerebrospinal fluid[30], and the elevation above healthy controls of a cryptic peptide in the gene HDGFL2 in C9orf72+ ALS CSF[13]. However, it is not clear that all cryptic splices are equally sensitive to TDP43 pathology, detectable, or have similar effects on gene or protein expression. Some may be very sensitive to loss of TDP43 activity, whereas others might require deeper TDP43 repression. Some cryptic splices may be located in non-coding exons, or result in unstable proteins, precluding peptide-level detection. Finally, even for an RNA or protein which stably accumulates in tissue, detection in accessible biosamples such as plasma or cerebrospinal fluid depends on transport to, and stability in, these compartments; properties which are very hard to predict. Therefore, having a complete catalog of cryptic splices will be essential to identify the optimal targets to establish robust assays.

Aside from purely physiological challenges, the potential universe of cryptic splices could depend on the cellular context, due to differences in expression of cryptic splice-harboring genes or the activity of splicing factors other than TDP43. Therefore, certain cryptic splices might be specific to spinal cord motor neurons lost in ALS, while others might reflect pathological processes in cortical neurons or other cell types, making them potentially useful in monitoring different subtypes of TDP43 pathology. Identifying context-specific cryptic splices has represented a particular challenge as most sequencing-based screens have focused on *in vitro* models exclusively [5, 14, 20] or a limited set of *in*

*vivo* data (primarily from ALS/FTD frontal cortex) [18]. These approaches, while allowing for a substantial enrichment of signal, may not be generalizable to the wider array of nervous system tissues affected by TDP43 pathology.

We therefore set out to create an expanded catalog of cryptic splices by meta-analyzing published RNAseq datasets from *in vitro* neuron models and post-mortem brain and spinal cord tissue from ALS and FTLD-TDP subjects, comprising 1,778 bulk RNAseq samples across 16 contexts. Our cryptic splice discovery pipeline [1, 10] identified 142 events across these datasets, including 76 which were not reported in previous analyses. Comparison of these cryptic splices across datasets revealed significant differences between cryptic splicing in cortical tissue, spinal cord, and *in vitro*. We validated a subset of these novel events with multiplex RT-qPCR on RNA extracted from spinal cords of controls and ALS subjects with histopathologically confirmed TDP43 pathology. Using several published single-nucleus RNAseq datasets of ALS and FTD cortex, we confirmed the elevation of cortical-enriched cryptic splices in disease, and localized them specifically to L2/3, L5, and L6 excitatory neurons.

# Results

*Post-mortem and in vitro transcriptomics datasets reveal diverse splicing patterns*

To identify cryptic splices associated with ALS, frontotemporal lobar degeneration with TDP43 pathology (FTLD-TDP), and experimental TDP43 depletion (TDP43-KD), we leveraged publicly available bulk RNAseq datasets (Table 1) comprising 1,871 samples (1,779 after QC filtering) of post-mortem patient RNA or sorted nuclei across 14 CNS regions and *in vitro* cell lines (Figure 1a) [5, 7, 14, 18, 24, 28, 32]. The post-mortem RNA samples included cerebellum, hippocampus, cortex and spinal cord from ALS, FTLD-TDP, FTLD-Tau, and control subjects generated by the New York Genome Center (NYGC) ALS consortium[5, 7, 28, 32] and, separately, the RiMOD-FTD consortium [24]. We also included a study comparing TDP43+ and TDP43- neuronal nuclei from ALS/FTLD-TDP frontal cortex (hereafter referred to as “Liu”) [18]. Finally, four datasets using siRNA-based TDP43-KD conditions in different neuronal models were included [5, 14].

We re-analyzed these datasets utilizing a single analytical framework, employing the *GSNAP* algorithm [34, 35] for read alignment followed by *SGSeq* and *DEXseq* R packages for splice detection and differential testing, respectively [1, 10]. We organized the 7 datasets into 16 “contexts” representing different tissue regions, data sources (e.g. NYGC and RiMOD-FTD frontal cortex were separate contexts), diseases, or different *in vitro* models, analyzing each context independently. Briefly, the SGSeq

framework generated a splice graph using ENSEMBL-annotated GRCh38 transcripts supplemented with novel exonic and intronic features observed from RNAseq alignments (Figure 1b). Variable splice graph regions with shared start and end nodes (*splice events*) were identified (Figure 1b) and the relative usage of paths through each event (*splice variants*) was quantified as the *variant frequency* (VF) for each sample, a number between 0 and 1 (analogous to the *percent spliced in* Ψ statistic reported by other methods). Across the 16 differential contrasts performed (see Methods), we calculated the difference in variant frequency (ΔVF) between the "TDP43-perturbed" (that is, ALS or FTLD-TDP disease, TDP43 knockdown, or TDP43- nuclei) and control conditions and observed as many as 431 significantly differentially spliced variants (DSVs: $q < 0.05$, $|\Delta VF| > 0.025$). Of these, up to 77 were "cryptic" splices, defined as those DSVs with mean or median VF < 0.05 in the control group (for example, see Figure 1c). The presence of cryptic splicing broadly agreed with regional expectations of TDP43 pathology: FTLD-TDP frontal and temporal cortex greater than ALS, ALS motor cortex greater than FTLD-TDP, and ALS spinal cord greater than non-spinal cord (Figure 1d). Cryptic splice events (shortened hereafter to "cryptic splices") were consolidated across contexts, yielding a catalog of 142 cryptic splices, consisting largely of cryptic exon inclusion events (40 cassette, 20 terminal, and 29 novel start exons), 31 novel exon skipping events, and a small number of retained introns and alternative splice sites (Figure 1e), with distinct tissue, disease, and model-specificity. 133/142 cryptic splices contained novel, unannotated splice junctions (6/9 cryptic splices with annotated junctions were from *in vitro* contexts). Our GSNAP/SGSeq analysis gave broadly concordant results to the STAR/MAJIQ-based approach previously published for NGN2-neurons [5] (Online Resource 1).

*Cryptic splicing exhibits spinal, cortical, and in vitro-enriched signatures*

To understand if the 142 identified cryptic splices showed context specificity across the included datasets we performed hierarchical clustering, which revealed six distinct clusters of enrichment representing disease, tissue, and model-specificity (Figure 2a, Online Resource 13). Distinct from these six splice sets, inclusion of the STMN2 cryptic exon 2a was unique in being context-agnostic, being elevated in ALS spinal cord, FTLD-TDP cortex, sorted TDP43- neurons, and all TDP43-KD models. The SGSeq-predicted cryptic junction coordinates of STMN2 aligned precisely with those reported previously (Figure 2b - left) [23]. We identified a spinal cord-enriched cryptic splice signature (7 splice variants), which was virtually absent from other CNS regions with TDP43 pathology or *in vitro* TDP43-KD models. One such splice was identified in CHCHD6, representing potential usage of an alternative first exon (Figure 2b - center). An *in vivo*-enriched signature (6 splice variants) was elevated across multiple post-mortem contexts but not in *in vitro* models. A cortical splice signature (24 splice variants) was

elevated in ALS-FTD/FTLD-TDP frontal/temporal cortex, ALS motor cortex, and *in vitro* models. This cortical signature included a novel terminal cryptic exon in HS6ST3 (Figure 2b - right), as well as the well-documented cryptic cassette exon in UNC13A (Online Resource 2)[20]. An *in vitro*-enriched signature (27 splice variants) was particularly evident in neuroblastoma cell lines, and largely absent *in vivo*. 39 splice variants formed an "FTD" signature, showing enrichment in FTLD-TDP and *in vitro* models, but not in post-mortem ALS. Lastly, a neuronal signature of 38 splice variants was enriched across TDP43-KD models and, partly, in TDP43-negative neuronal nuclei from FTLD-TDP frontal cortex. Examples of splices from each signature are highlighted in Online Resource 2.

Comparison of differential splicing results across contexts supported this clustering-based categorization. Spinal cord subregions showed concordant elevations of spinal-enriched splices in ALS, which was not observed in the FTLD-TDP cortex, and vice-versa (Online Resource 3). Cortical-enriched splices were also perturbed to a greater extent than splices in the FTD splice set (Online Resource 4). Interestingly, *in vitro* models showed greater similarity to FTLD-TDP cryptic splicing than to ALS motor cortex or spinal cord (Online Resource 4).

Given our definition of "cryptic" (mean or median VF < 0.05 in control group) is also context-specific, we considered whether any cryptic events might be non-cryptic (VF ≥ 0.05 in controls) in another context. We found that 16/142 splices exhibited broad inclusion across *in vivo* datasets (mean VF of cryptic splice > 0.1) regardless of diagnostic group (Online Resource 5). Critically, all such splices were discovered from *in vitro* analyses, and 11/16 were members of *in vitro* splice sets, with 5 in the FTD splice set (which is also enriched in *in vitro* models). We omitted these splices in further downstream analyses, such as the generation of splice scores, but retain them here given they reflect cryptic splicing within a particular *in vitro* context. Importantly, no cryptic splices derived from *in vivo* analyses were non-cryptic *in vitro*.

*Context-specific gene expression may explain cortical cryptic splicing, but not other signatures*

We next explored potential mechanisms for context-specific splicing. One potential explanation could simply be context-specific expression of the genes which harbor cryptic splices ("host genes"). We determined the fold-change in expression of host genes within each cryptic splicing signature (Figure 3a), in control conditions, across datasets mirroring the enrichment observed in the splicing signature. For example, *in vitro* signature host genes were compared between *in vivo* and *in vitro* datasets, to determine if their elevated expression in *in vitro* contexts might explain their context-specific cryptic splicing (see Figure 3 caption). We then computed the percentile of the fold-changes relative to the entire genome to determine the relevance of these differences. No cryptic splice host gene showed true specificity (i.e.

being unexpressed outside their expected context), though 25/136 genes were above the 95th percentile of their respective genome-wide LFC distribution (Figure 3b). These genes were largely (18/25) from the cortical and FTD splice sets, indicating that greater cortical expression of host genes for these two signatures could explain some of their context-specific splicing. The remaining signatures showed either minimal, or even inverse (spinal), context-enrichment of host genes. Therefore, gene expression differences of host genes across contexts may explain some cortical-enriched cryptic splicing, but not other context specific cryptic splicing signatures.

*Validation of cryptic splices by RT-qPCR and RNAseq*

To confirm our bioinformatically-predicted cryptic splices, we developed RT-qPCR assays targeting the cryptic and canonical variants of 24 cryptic splice events (Online Resource 14) and quantified their expression in an independent cohort of ALS and healthy control lumbar spinal cord sections (Table 2). Of these 24, we predicted three cryptic splices would be elevated based on our RNAseq-based meta analysis: STMN2 as well as two spinal-specific events which, to our knowledge, have not been previously reported, in CHCHD6 and COL22A1. We observed significant elevations of these splices, as well as five additional cortical and FTD splices (in HDGFL2, HS6ST3, ATXN1, CYFIP2, and TRRAP) though at more modest significance levels (Figure 4a). Cortical and FTD splices which did not show elevations were unamplified in over 40% of samples, suggesting that most cortical and FTD splices are either minimally expressed or undetectable in the spinal cord (Figure 4b).

However, to understand why some assays targeting “non-spinal” events, such as those in HS6ST3 and HDGFL2, nonetheless did show ALS-enrichment in the spinal cord, we performed RNAseq on the same cohort. We observed significantly increased inclusion of the STMN2 and CHCHD6 cryptic splices in ALS, which correlated with RT-qPCR results (Figure 4e), but other cryptic splices detected in RT-qPCR were not detected in RNAseq. Therefore RT-qPCR is more sensitive than typical bulk RNA sequencing to detect cryptic splicing. Our splice sets, defined based on RNAseq analysis, therefore may be context-*enriched* rather than context-*specific*.

We also sought to confirm the association of these cryptic splices with TDP43 pathology via pTDP43 S409/410 immunohistochemistry (IHC) in tissue sections neighboring those used for molecular assays (Figure 4c). pTDP43 staining was classified into 4 categories ranging from 0 (no pTDP3 aggregates) to 3 (many cells showing abundant pTDP43 aggregates), by a neuropathologist blinded to the donor diagnoses (see methods, Figure 4d). pTDP43 scores were significantly correlated with STMN2 and CHCHD6 cryptic splice variant frequency in RNAseq (Figure 4f). Moreover, all cryptic splices showing ALS-enrichment in RT-qPCR (with the exception of CYFIP2 and TRRAP, and addition of UNC13A)

were significantly positively correlated with pTDP43 scores (Figure 4g). Overall, these IHC, RT-qPCR, and RNAseq experiments validate a number of novel and extant cryptic splices in ALS spinal cord, demonstrate the superiority of RT-qPCR relative to standard RNA-Seq for detecting cryptic splices, and show the association of these cryptic splices with underlying disease biology.

*snRNAseq localizes cryptic splicing in cortex to L2/3, L5, and L6 IT excitatory neurons*

Finally, we sought to determine the cell-type enrichment of our cryptic splices, utilizing snRNAseq datasets from two recent studies profiling (1) motor cortex from sporadic and C9orf72 ALS and FTLD patients (Pineda et al.) [27] and (2) motor and frontal cortex (analyzed separately) from C9orf72 ALS and FTLD patients (Li et al.) [15]. After QC and mutual-nearest-neighbor batch correction [12], 415,741 nuclei were retained from Pineda yielding 29 clusters (Figure 5a). Likewise, analysis of Li motor cortex yielded 59,610 cells forming 31 clusters (Online Resource 8A), while the frontal cortex dataset comprised 63,076 nuclei forming 34 clusters (Online Resource 10a).

In both datasets, we quantified cryptic splice junctions from CellRanger filtered BAM alignments, 41 cryptic splices were detected in Pineda (Online Resource 6: upper), with six at ≥10 reads across the whole dataset (Figure 5b). These were all terminal or penultimate cryptic exons, reflecting the inherent 3' bias in snRNAseq data. Detection of cryptic splice junctions was greater in Pineda than in either cortical region of Li, likely due to the 7-fold greater number of nuclei. Notably, these six cryptic splices were amongst the most abundant quantified in both the motor and frontal subsets of Li (Online Resource 6: middle and lower), suggesting their predominant detection is not artifactual.

ARHGAP32, STMN2, and HS6ST3 cryptic splices were largely localized to L2/3, L5, and L6 excitatory neurons (Figure 5b, Online Resources 8b and 10b), and were virtually undetected in controls (Figure 5c, Online Resources 8c and 10c). Differential abundance analysis of cryptic-junction-harboring cells using edgeR, in Pineda, confirmed this specificity as statistically significant for ARHGAP32 in L2/3 (in C9-ALS, sFTLD, and C9-FTLD), L5 (C9-ALS only) and L6 excitatory neurons (in sFTLD only), for STMN2 in L2/3 neurons (in sFTLD), and for HS6ST3 (Figure 5c). Other *in vivo*, cortical, and FTD splices, in aggregate, were detected in a virtually disease-specific manner (though not statistically significant), while *in vitro*-enriched splices were not disease-specific and were also not significant (Figure 5c). Non-neuronal cell-types showed sporadic and minimal inclusion of cryptic splices (Online Resources 7, 9, and 11)

Analysis of Li motor and frontal cortex was similar to Pineda, though only elevated STMN2 (in C9-FTLD) cryptic splicing in L6 excitatory neurons was significant, likely reflecting the greater sparsity of this dataset (Online Resources 8c and 10c). Apart from STMN2, C9-ALS samples had greater

detection of cryptic splices than C9-FTLD, representing the main contrast from the Pineda analysis. ARHGAP32, HS6ST3, and other *in vivo*-enriched splices were observed virtually disease-specifically (though again not statistically significant) in L2/3 excitatory neurons in C9-ALS motor cortex.

This analysis validates ARHGAP32 and HS6ST3 as novel cryptic splicing targets and localizes cryptic splicing to neurons, possibly with enrichment in intra-telencephalic projection neurons of layers 2/3, 5 and 6. The failure to detect cryptic splices in the L5 ET neurons, which would include the Betz cells expected to have TDP43 pathology in ALS, may be due to the combination of very small numbers of such cells, rarity of cryptic junctions in these datasets, and potential survivor bias in remaining Betz cells in ALS (i.e. cells with the most TDP43 pathology may already have degenerated pre-mortem). Motor cortex cryptic splicing (reaching statistical significance) appeared more prevalent in spontaneous-FTLD than ALS, if modestly. While this difference cannot be definitively confirmed here (particularly given the sparsity of snRNAseq data), it is consistent with modestly (though significantly) elevated HS6ST3, STMN2, and ARHGAP32 cryptic splice inclusion in ALS/FTD vs ALS bulk motor cortex RNAseq (Online Resource 12).

# Discussion

In this study, we meta-analyzed nearly 1,800 published RNAseq profiles of ALS and FTLD-TDP post-mortem tissue, and *in vitro* TDP43-KD models, identifying 142 cryptic splices, 76 of which had not been identified in previous cryptic exon screens. We described distinct splicing signatures across *in vivo* and *in vitro* contexts, most clearly in the form of spinal and cortical signatures which were predominantly observed in ALS and FTLD-TDP, respectively. We validated eight of these splices in an independent cohort of ALS lumbar spinal cord samples, where we observed an enrichment of the spinal splicing signature. Lastly, investigation of snRNAseq profiles of motor and frontal cortex demonstrated that cryptic splicing at the 3' end of transcripts is detectable at single-nucleus resolution, and disease-enriched in L2/3, L5, and L6 excitatory neurons.

Our primary finding in this study was the discovery that cryptic splicing profiles of TDP43-proteinopathies are context-dependent, with particularly notable differences in spinal and cortical cryptic splicing signatures. In RNAseq these signatures were largely distinct, and appeared to overlap only with STMN2. RT-qPCR experiments, with superior limits of detection to RNAseq, revealed that some cortical cryptic splices are included in ALS spinal cord at levels, on average, many-fold lower than STMN2 and spinal splices. Additionally, we observed a divergence in cryptic splicing in which a number of splices were observed exclusively *in vivo*. *In vitro* cell-lines are imperfect models of ALS and

FTLD-TDP biology, though the degree to which TDP43-KD models failed to replicate a number of highly significant spinal cryptic splices (e.g. CHCHD6 and COL22A1) was nonetheless striking. Our findings particularly highlight the limitations of existing *in vitro* TDP43-KD models for investigating ALS spinal-related cryptic splicing, where STMN2 was the only common significantly elevated cryptic splice, even in an embryonic-derived motor neuron TDP43-KD model (Online Resource 4c,d) [14]. The cortical splice signature, in contrast, was elevated in both ALS motor cortex and FTLD-TDP frontal and temporal cortex, albeit more robust in the FTLD-TDP cortex, and also showed a broad overlap with *in vitro* models (Online Resource 4b). This cortical signature thus represents the main locus of similarity across cortical ALS, FTLD-TDP, and *in vitro* models, with the FTD signature representing further similarity of FTLD-TDP and *in vitro* models, and the other signatures representing how spinal, cortical, and *in vitro* cryptic splicing differs.

The preponderance of novel cryptic splices identified in this study reflects our investigation of *in vivo* tissue, as our analyses of *in vitro* models were concordant with those previously reported (Online Resource 1). Cryptic splicing of STMN2 and novel cryptic splices in ARHGAP32 and HS6ST3, were disease-enriched in our investigation of snRNAseq datasets, particularly in cellular populations known to be enriched in TDP43 aggregates (Figure 5, Online Resources 8 and 10). Similar to another recent investigation of cryptic splicing in ALS/FTD frontal cortex, cryptic splices were sparse in snRNAseq but nonetheless detectable and enriched in disease [9]. Our localization of elevated cryptic splices in L2/3, L5, and L6 IT neurons broadly replicates this recent study by Gittings and colleagues who also observed the greatest STMN2 cryptic splicing in similar layer-specific IT neuron populations [9]. TDP43 inclusions are largely localized in L2/3 of FTLD-TDP neocortex and motor cortex, extending to L6 in non-type A FTLD-TDP, consistent with our localization of FTLD cryptic splicing in L2/3, L5, and L6 IT neurons [25, 26]. Cryptic splices in ALS were localized to L2/3 and L5 IT neurons (though disease-elevated with less statistical significance given the sparsity described above), which have also been reported to harbor TDP43-pathology in addition to the well-characterized pathology in Betz cells [4].

Notably, tissue or cell-type-specific expression of cryptic splice host genes was only a modest driver of tissue-specific cryptic splicing (Figure 3). Cryptic splices in the cortical signature (and FTD signature to a lesser degree) were largely in cortical-enriched genes (Figure 3a), suggesting that this cortical-enriched cryptic splicing could be by way of context-enriched host gene expression. Other splice sets showed only a small number of cryptic splices with enriched host gene expression that could explain their context specificity. Another possible explanation for these differences could be tissue- or cell-specific ensembles of RNA binding proteins capable of compensating for loss of TDP43 activity at different cryptic exons. Integration of these and other datasets (such as CLIP-Seq atlases) could generate mechanistic hypotheses which could be tested *in vitro*.

Further studies should also aim to address the limitations of our current work. First, we did not systematically validate the TDP43-dependence of our cryptic splices, beyond those correlated with pTDP43 IHC measures. While we found them associated with several metrics reflective of TDP43 pathology (e.g.. diagnosis, TDP43-pathology enrichment, pTDP43 IHC), experimental confirmation in improved TDP43-KD models, mapping TDP43 binding sites, and determining rescue by steric inhibition of these sites will provide greater confidence in their mechanistic relation to TDP43 aggregation. Nonetheless, the identified splices have potential for ALS/FTLD-TDP biomarker development, given their discriminatory power. Second, while bulk-RNAseq enabled a large-scale investigation of cryptic splicing, direct comparison with RT-qPCR revealed limit-of-detection for some cryptic splices. Future studies should endeavor to increase sequencing depth and prioritize tissue with high RNA quality to improve detection of rare cryptic splices, and utilize long-read sequencing technologies to confirm full cryptic transcript isoforms. Lastly, studies utilizing cellular purification methods, such as flow cytometry, to isolate cells with evidence of TDP43 pathology should be conducted in the spinal cord and motor cortex (in ALS and FTLD-TDP) for a deeper characterization of cryptic splicing in these contexts.

In summary, our meta-analysis revealed novel cryptic splices and hitherto uncharacterized tissue, disease, and model-specific cryptic splicing signatures related to TDP43 proteinopathies. The profound differences in cryptic splicing between the spinal cord and cortex, ALS and FTLD-TDP, and *in vivo* and *in vitro* contexts have important implications for disease biology and biomarker development.

# Methods

*Data ingestion, QC, and raw data analysis*

Publicly-available RNAseq datasets were downloaded from the Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), or European Genome-Phenome Archive (EGA), as available. Dataset sources and descriptions are listed in Table 1. In all cases, raw, unprocessed FASTQ files were obtained for use with our GSNAP/HTSeqGenie-based alignment and quantification pipeline [11, 34, 35]. All FASTQ files were processed with FastQC [36], and files failing the *Basic Statistics* or *Per base sequence quality* measures were excluded from downstream analysis.

Bulk RNAseq data was processed using HTSeqGenie [11]. First, reads with low nucleotide quality (70% of bases with Phred quality <23) or matches to rRNA and adapter sequences were removed. The remaining reads were then aligned to the human reference genome (GRCh38.p10) using GSNAP [34, 35] version ‘2013-10-10-v2’, allowing maximum of two mismatches per 75 base sequence (parameters:

‘-M 2 -n 10 -B 2 -i 1 -N 1 -w 200000 -E 1 --pairmax-rna=200000 --clip-overlap'). Transcript annotation was based on the Gencode genes database (GENCODE 27). To quantify gene expression levels, the number of reads (or concordant fragments for paired-end chemistry) mapping unambiguously to the exons of each gene was calculated.

snRNAseq data was processed using CellRanger v7.1.0 using largely default parameters. Reads were aligned to the human reference genome GRCh38, and gene expression quantification was performed using *cellranger count* (“--chemistry” was set to “threeprime” and “–include_introns” was set to true).

*Splice discovery and quantification with SGSeq*

We used the SGSeq R/Bioconductor package [10] to discover and quantify alternative splicing. GRCh38 reference annotations were used to build a splice graph describing the potential splicing of each gene, in which non-overlapping exon bins are represented as nodes and splice junctions as edges (`SGSeq::convertToTxFeatures`). Additional splice graph features, not present in the reference annotations, were identified from the input BAM alignments (`SGSeq::predictTxFeatures`). Regions of the splice graph with multiple paths and shared start and/or end nodes represent *splice events*, with the particular paths representing *splice variants* (`SGSeq::findSGVariants`). Junction reads supporting the 5’ and 3’ ends of splice variants were quantified. The relative usage of each splice variant, *variant frequency* (VF), was calculated as a ratio of the junction reads supporting the start or end of the variant versus the sum of such reads for all variants in a splice event (`SGSeq:::getSGVariantCountsFromBamFiles`; e.g. figure 1B). For cases such as cassette exons with multiple junctions, the weighted mean of the variant start and end were used.

SGSeq analyses were performed separately within datasets to improve cryptic splice detection across varying contexts, with four combined analyses performed across similar contexts to increase sample sizes (indicated in Online Resource 14). This reflected combining all spinal cord, NYGC frontal/temporal cortex, and *in vitro* contexts into respective SGSeq analyses. Parameters controlling the inclusion of splice graph features were modified slightly to optimize detection across datasets with differing sequencing depths and levels of intronic reads. Namely $\beta$ (the FPKM threshold for splice junction inclusion) and the *min_samp* parameter (minimum number of samples in which a splice graph feature must be observed to be included) were modified as described in Online Resource 14. Splice variant nomenclature has the format: “GeneID_EventNumber_VariantNumber/TotalVariants_EventType” (e.g. STMN2_2_1/2_AE).

*Differential splicing analyses with DEXseq*

Variant count quantifications were used as input to DEXseq to test for differential usage of splice variants within an event, similar to a previously described approach [31]. Splice variants were included if supported by at least 10 reads, in at least *n* samples (n=2 for in vitro models, n=4 for Liu and RiMOD, and n=8 for all other analyses). Any events with a single variant remaining (i.e. singletons) were excluded from analysis, as they were considered effectively constitutively spliced. After filtering, DEXseq analysis proceeded as previously described [31]: normalizing for sequencing depth (`DEXSeq::estimateSizeFactors`), estimating transcriptome-wide dispersion (`DEXSeq::estimateDispersions`), and testing for differential variant usage within an event (`DEXSeq::testForDEU`).

Differential contrasts were performed between ALS and controls in the hippocampus, cerebellum, spinal cord, motor cortex, frontal cortex, and temporal cortex, between FTLD-TDP and controls in the frontal cortex, temporal cortex, and RiMOD datasets, between TDP43-negative nuclei and TDP43-positive nuclei from the Liu dataset, and between TDP43-KD siRNA treatments and scrambled siRNA controls in *in vitro* contexts. Statistical significance was set at a Benjamini-Hochberg[3] FDR of 5% and a ΔVF of 0.025. Cryptic splices were identified as significantly differentially spliced variants with a control group mean or median VF of 0.05 or less.

Variant frequencies within an event sum to 1, thus inclusion of all variants in DEXseq analysis would result in excessive multiple testing correction. In our DEXseq analysis, we excluded the first variant from each event to address this, despite this resulting in some disease-elevated cryptic variants being excluded (though significance is identical for the canonical variant with 1-VF). For all significantly differentially spliced variants, the disease-elevated variants were confirmed *post-hoc*. Cryptic splice variants were consolidated across all analyses by de-duplicating independently identified variants with identical cryptic junctions.

Variant frequency Z-scaling (used only in visualization of the figure 2A heatmap) was used to more clearly compare variant frequency across datasets with vast differences in effect size. Variant frequency was normalized to the mean and standard deviation of the respective control group, with a pseudocount included for the standard deviation to avoid dividing by zero when a splice was completely undetected in controls in a particular context. $VF_{Z-norm} = \frac{VF\ for\ sample - \overline{X}(Control\ VF)}{\sigma(Control\ VF) + 0.01}$

*Post-mortem lumbar spinal cord tissue*

All human tissue samples were supplied by Discovery Life Sciences and Fidelis Research. Each supplier received IRB approval of research, appropriate informed consent of all subjects contributing biological

materials, and all other authorizations, consents, or permissions as necessary for the transfer and use of the biological materials for research at Genentech.

*pTDP43 IHC, imaging, and quantification*

Frozen spinal cord samples embedded in OCT were sectioned transversely at 10 um thickness, and IHC was performed using the Ventana Discovery Ultra platform (Ventana Medical Systems). Sections were fixed for 30 minutes in neutral buffered formalin, heat antigen retrieved with Ventana CC1 for 16 minutes (cat. no. 950-500, Ventana Medical Systems), blocked for 4 min (DISCOVERY Inhibitor; Ventana, 760-4840), and incubated with 1:2000 mouse anti-human TDP43 pS409/410 (Cosmo Bio TIP-PTD-M01A; clone 11-9). Signal was detected with Ventana Mouse Omnimap anti-mouse HRP for 16 minutes, and labeled with Ventana Discovery Purple incubation for 12 minutes. The sections were counterstained with Hematoxylin II (Ventana Medical Systems; Tucson, AZ), dehydrated, and cover-slipped. Negative control IHC was performed using naive mouse IgG1 antibody.

Semiquantitative analysis of pTDP43 IHC was performed by a pathologist (JH) blinded to the cryptic splicing quantifications and donor diagnoses, and was scored as follows: 0, No pTDP43; 1, Minimal pTDP43 aggregates in rare cells; 2, Moderate pTDP43 aggregates in scattered cells; 3, Severe pTDP43 aggregates in frequent cells. No healthy control samples had pTDP43 aggregates in the sections examined.

*RT-qPCR validation of cryptic splices*

Spinal cord tissue was homogenized using Bertin-Precellys homogenizer in Buffer RLT (Qiagen) supplemented with beta-mercaptoethanol per manufacturer recommendations. RNA was prepared using the RNeasy Plus RNA isolation kit (Qiagen) and quantified using Tapestation High sensitivity RNA kit (Agilent). Levels of RNA were normalized for cDNA synthesis with SuperScript IV VILO master mix (Thermo Fisher). Transcripts were quantified by real-time quantitative PCR on a Quantstudio 7 Real-Time PCR instrument (Thermo Fisher). Probes for 24 cryptic splice variants (and the respective canonical variants) were designed using Primer3 (Online Resource 15) [33], and synthesized by Integrated DNA Technologies. For each sample, mRNA abundance was normalized to GAPDH (Hs02786624_g1, Thermo) and HPRT1 (Hs02800695_m1,Thermo) transcripts. Triplicate PCRs were performed on each sample of cDNA and the mean value of each sample was used in statistical analysis.

We quantified relative splice variant expression using the ΔΔCt method:

$\Delta\Delta Ct = \Delta Ct\ (Sample) - \overline{X}(Control\ \Delta Ct)$ where:

$\Delta Ct = Ct(Assay\ of\ interest) - \overline{X}(Ct_{HPRT1} + Ct_{GAPDH})$. For ΔΔCt values in Figure 4A, cycle thresholds (Cts) for unamplified assays were imputed as $(1 + Largest\ Ct\ for\ assay)$, no imputation was performed for ΔCt values (i.e. not normalized to experimental controls) shown in any other figure. ΔΔCt values were compared between ALS and control samples by unpaired t-test, significance was set at $p < 0.05$.

*Bulk-RNAseq of lumbar spinal cord samples from ALS and controls*

Total RNA was quantified with Qubit RNA HS Assay Kit (Thermo Fisher Scientific) and quality was assessed using RNA ScreenTape on TapeStation 4200 (Agilent Technologies). For sequencing library generation, the Truseq Stranded mRNA kit (Illumina) was used with an input of 100-1000 nanograms of total RNA. Libraries were quantified with Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and the average library size was determined using D1000 ScreenTape on TapeStation 4200 (Agilent Technologies). Libraries were pooled and sequenced on the NovaSeq 6000 platform (Illumina) to generate 50 million paired-end 100bp reads for each sample. Alignment and splice variant quantification were performed as described above.

*snRNAseq QC, analysis, and cell-type annotation*

snRNA-seq data from Pineda *et al* (Motor Cortex), Li *et al* (Frontal Cortex), and Li *et al* (Motor Cortex) were analyzed separately, using the scuttle, scran, and scater packages [15, 19, 21, 27]. snRNA-seq quality control metrics (total number of unique molecular identifiers (UMIs), number of detected genes, and mitochondrial read percentage) were computed for each nucleus, and sample-wise outlier-based filtering was performed to remove nuclei with any metric >3 median absolute deviations from the sample median. After initial analysis, both cortical regions of the Li dataset contained a cluster of low-UMI, high mitochondrial UMI percentage, poorly annotated nuclei which were filtered by removing nuclei with fewer than 1800 UMIs or greater than 10% mitochondrial UMIs (followed by re-analysis). UMI counts were size-factor normalized within each sequencing batch, scaled across batches to match the batch with the lowest coverage, then log-transformed. Principal component analysis was performed on the 2000 most variable genes within each dataset. Batch correction was performed through mutual nearest-neighbor (MNN) integration across donors (Pineda) or biological sex (Li) on 30 components [12]. 25

MNN-corrected components were used as input to shared nearest-neighbor clustering, and further dimension reduction by uniform manifold approximation and projection (UMAP).

Cell-type annotation was performed using SingleR[2] and the Allen Brain human M1 cell-type atlas (https://portal.brain-map.org/atlases-and-data/RNAseq/human-m1-10x) as a reference. SingleR annotations were confirmed by cross-referencing with population enrichment of canonical cell-type and laminar markers: SLC17A7 (excitatory neurons), CUX2, RORB, THEMIS, FEZF2 (superficial to deep layers), GAD1, GAD2, SST, PVALB, VIP (interneurons), MBP, MAG (oligodendrocytes), PDGFRA (oligodendrocyte precursor cells), TMEM117, IBA1 (microglia), GFAP, AQP4 (astrocytes). Doublet detection was performed using scDblFInder [8], and clusters enriched in high doublet scores which exhibited expression of cell-type marker genes from different cell-types were removed from further analysis.

*Splice junction quantification and differential abundance*

To identify cryptic splice junctions within single-nuclei, we used the Bioconductor GenomicAlignments and Rsamtools package to extract reads from CellRanger BAM alignments with split alignments covering cryptic splice junctions using the `Rsamtools::ScanBamParam` and `GenomicAlignments::readGAlignments` functions, for each splice. Reads were then filtered to confirm that only those compatible with the cryptic splice junction, and containing coverage in flanking cryptic or canonical exons, were included. Cryptic junction read counts were then mapped to the corresponding nucleus via sample identifiers and cell barcode tags. Given the low dynamic range in junction counts per nucleus (Online Resource 6), downstream analyses focused on the sample-wise proportion of cryptic splice-positive nuclei. For differential abundance analyses, cryptic splice-harboring cells were categorized as an additional cell-type (e.g. “cryptic STMN2 L2/3 neurons”) and included with non-cryptic cell-types. Sample-wise abundances of each cell-type were then compared across diagnostic groups and corrected for multiple comparisons using edgeR [6, 29]. Significance was set at $q < 0.05$.

*Data Availability*

Previously published bulk and snRNAseq data sources are available in Table 1. We have deposited the raw bulk RNAseq data from our new cohort of 10 control and 10 ALS lumbar spinal cord samples to GEO (GSE255683). Cryptic splice variant coordinates and quantifications are available in Online Resource 13, pTDP43 IHC data is available in Table 2, and RT-qPCR data is available in Online Resource 16.

Figure Captions

**Figure 1. Analytical workflow used to meta-analyze cryptic splicing in post-mortem and *in vitro* RNAseq**. **a** - Data sources and respective sample sizes by experimental group. Named tissues (Cerebellum through Temporal) are all from the NYGC ALS Consortium; RiMOD consortium data are from frontal cortex. **b** - Conceptual framework used in SGSeq to identify splicing events (left) and test for differential splicing using DEXseq (right). Known transcript models and novel exonic and splice junction features identified from input RNAseq data are used to build a splice graph describing the potential splicing of a given gene. Splice events are then identified, describing regions of the splice graph with multiple potential paths and sharing start and/or end nodes. Three example events are shown representing (1) an alternative first exon, (2) a cassette exon (with a cryptic exon inclusion), and (3) an alternative last exon. Within a splice event, each path is termed a splice variant (here, identified using small numbers annotating features), the inclusion of which is quantified by dividing the number of junction reads supporting the 5' and 3' ends of the variant by the total reads supporting that splice event (variant frequency). DEXseq is then used to identify splice variants with significant differences in variant reads across contrasts of interest. **c** - Representative examples of our categorization of splicing, specifically the distinction between cryptic and non-cryptic differentially spliced variants. Grey dashed lines indicate control variant frequency threshold of 0.05. **Left**: no significant difference in the canonical glia-enriched skipping of APP exon 8. **Center**: significantly greater inclusion of the STMN2 cryptic exon (note the low control variant frequency). **Right**: significantly greater inclusion of a non-cryptic splice in ADD1 (note the control variant frequency above 0.05). Events and variant numbers correspond, broadly, to 5' to 3' order along gene length. **d** - Number of cryptic splice variants identified in each differential contrast performed in context-specific analyses. CNS region/dataset/model is indicated on the y-axis, and differential contrast indicated on subplot titles. **e** - Overview of cryptic splices identified in this study, across all contrasts, categorized by splice type. Splices categorized as previously-reported have been described in one or more of Brown et al (2022), Ma et al (2022), Klim et al (2021), or Seddighi et al (2023).

**Figure 2. Cryptic splicing shows context-specificity, particularly between cortex and spinal cord. a** - Heatmap showing inclusion of 142 cryptic splices across all datasets included in this study. Variant frequency was Z-normalized (see methods) to the mean and standard deviation of controls of each context for scaling purposes, given the wide range of variant frequencies represented. Column annotation bars represent diagnostic groups and dataset sources. Control groups are colored in grey, and contexts with TDP43 depletion confirmed by experimental condition, *post mortem* histopathology, or presence of genetic factors are colored in gold. **b** - Variant frequency and coverage plots of three illustrative cryptic splices in STMN2 (left), CHCHD6 (centre), and HS6ST3 (right). Variant frequency plots (above legend) contain the sample-wise quantification of the cryptic variant denoted in the coverage plots.*** indicates a DEXseq FDR-corrected q-value < $10^{-6}$. Next, "Mean Coverage" track shows mean RNAseq coverage, aggregated by group and tissue, in histograms and splice junctions with arrows. Junction coordinates are indicated by arrow start and end positions, with the height indicating the junction read depth, and strand indicated by position of arrow head (minus strand left, plus strand right). For visual clarity, junctions with less than 10 reads across all samples, and those on the opposite strand as the gene of interest, are not shown. Cryptic exon coordinates are highlighted in green. "Ensembl Transcripts" track shows canonical

Ensembl transcripts present in the same coordinate space (rectangles: exons, arrows: introns, with arrows pointing from splice donor to acceptor site). Bottom track shows the cryptic and canonical splice events identified by SGSeq, in which arrows indicate splice junctions and boxes indicate exonic features. Note the spinal cord-specificity of the CHCHD6 cryptic splice, cortical-specificity of the HS6ST3 cryptic splice, and the presence of the STMN2 cryptic splice in both tissues.

**Figure 3. Context-specific cryptic splicing is modestly related to context-specific host gene expression.** Cross-context cryptic splice host gene expression differences compared to genome-wide distributions. **a**: $Log_2$ fold-change (LFC) in expression of cryptic splice host genes ("cryptic") within each splice set and all other genes ("genome"), in control subjects of observed contexts of enrichment compared to controls of non-enriched contexts. Expected enrichment, based on splice set enrichment in Figure 2A, and LFC comparisons used to generate distributions are illustrated above each splice set. STMN2 was included as a cryptic splice in each set for reference only. Dashed line indicates how LFC percentiles were determined using spinal splice COL22A1 as an example. The dashed line indicates that the inter-context LFC of COL22A1 (1.86) corresponded to the 93.1 percentile of all inter-context LFCs. **b**: LFC percentile for each cryptic splice, relative to the genome-wide LFC distribution for each comparison of contexts. Solid horizontal line indicates 50th percentile and dashed lines indicate 5th and 95th percentiles. Genes with LFCs beyond the 5th and 95th percentiles, and those significantly elevated in qPCR experiments (including STMN2), were labeled.

**Figure 4: Validation of cryptic splicing in the lumbar spinal cord of an independent ALS cohort. a** - Eight cryptic splices showing significant elevations (by unpaired t-test) in an independent cohort of 10 ALS (blue) and 10 healthy control (grey) lumbar spinal cord samples. -ΔΔCt values were calculated relative to the mean of HPRT1 and GAPDH loading controls, and then normalized to the mean of control samples (see methods). Novel cryptic splices (same criteria as described in figure 1e) are highlighted in green. ALS and control groups were compared using unpaired t-tests * $p < 0.05$, ***: $p <= 0.001$, ****: $p <= 10^{-4}$. **b** -ΔCt values (i.e. normalized to loading controls but not experimental controls) for all 24 cryptic splice assays, grouped by *a priori* validation expectation and splice set. Horizontal bars indicate mean -ΔCt values for ALS (blue) and control (grey) samples. Percentage of assays which were unamplified for ALS and controls are indicated on each subplot. **c** - Representative images of pTDP43 IHC in a control (upper panel) and ALS (lower panel) lumbar spinal cord section. Anatomical landmarks (AH - anterior horn; dashed line - grey/white interface; all images oriented with anterior side at the bottom) are noted on low-magnification images. Boxed regions of interest are highlighted in high-magnification images (right), showing strong non-nuclear intense staining of pTDP43 in ALS motor neurons (black arrows), and only diffuse background stain in control motor neurons (white arrows). Scale bars: 500 um (large images), 25 um (small images). **d** - pTDP43 IHC scores (semi-quantitative) for control and ALS spinal cord sections. Note that 2 ALS and 1 control section had uninterpretable IHC (thus are NA values in this and subsequent plots). **e** - Correlation of qPCR and RNAseq quantifications of STMN2 and CHCHD6 cryptic splices (i.e. those which were detected in both qPCR and RNAseq). Significant Pearson correlation coefficients are labeled on each plot. *$p< 0.05$. **f** - Correlation of STMN2 and CHCHD6 variant frequency (RNAseq) with pTDP43 scores, Spearman's ρ correlation coefficients are visualized on each plot. *$p < 0.05$. **g** - Correlation of -ΔCt values (normalized to GAPDH and HPRT1) for each cryptic splice, with pTDP43

scores. The top row contains cryptic splices containing those with significant elevations (i.e. 4A). Spearman's ρ correlation coefficients are visualized on plots with significant correlations. *$p < 0.05$.

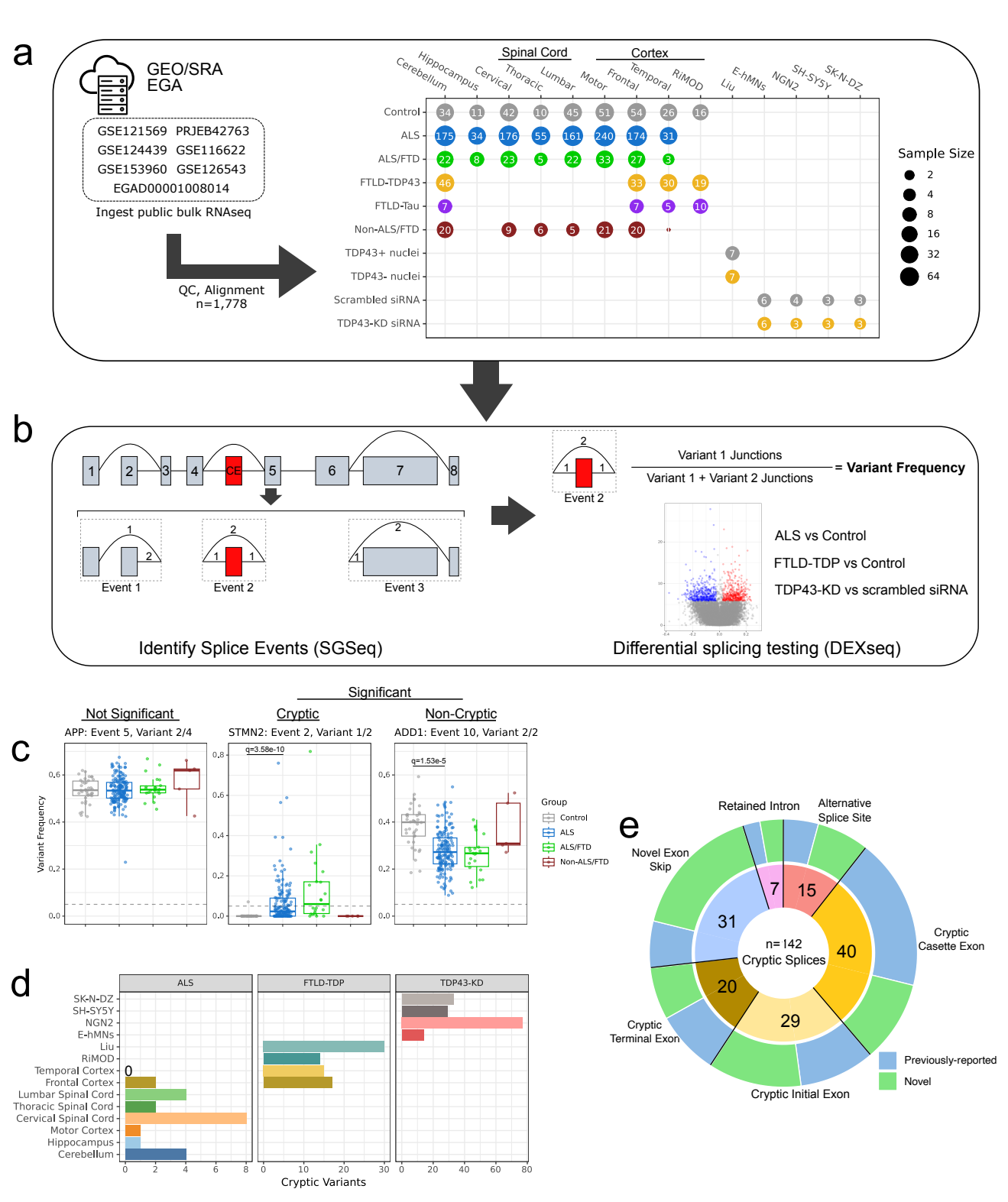
**Figure 5. Cryptic splices are enriched in L2/3, L5, and L6 excitatory neurons in ALS and FTLD motor cortex in snRNAseq. a -** Uniform Manifold Approximation and Projection (UMAP) embedding showing identified cell populations in analysis of snRNAseq data from Pineda *et al*. **b -** UMAPs showing cryptic splice-harboring cells in yellow, for splices detected at ≥10 junction reads. Dashed circle highlights populations of excitatory neurons harboring the majority of cryptic splicing. **c -** Cellularity plots showing the percentage of cryptic-positive cells, for each cell-type, relative to the total number of detected cells (per donor). "Other excitatory" combined L5 ET, L5/6 NP, L6b, and L6 CT populations, Interneurons combined Lamp5, VIP, SST, and PVALB populations, with all other populations combined in "Other Cell Types". Population-wise cryptic cellularity for "Other Cell Types" is visualized in Supplementary Figure 7. Each dot represents an individual donor. Significance was derived from differential abundance testing using edgeR, to account for potential composition effects. * $q<0.05$, ** $q<0.01$, *** $q<0.001$.

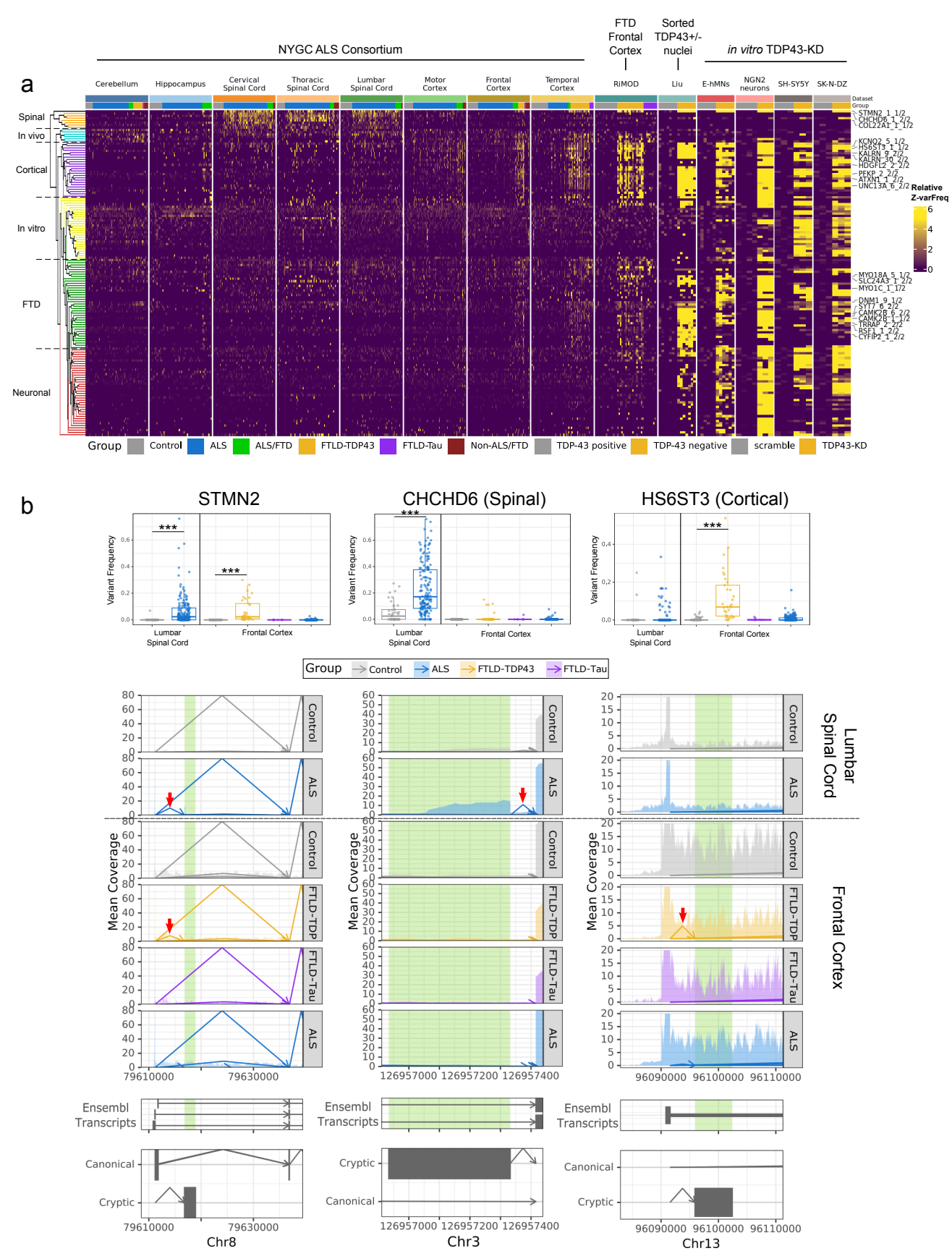
# References

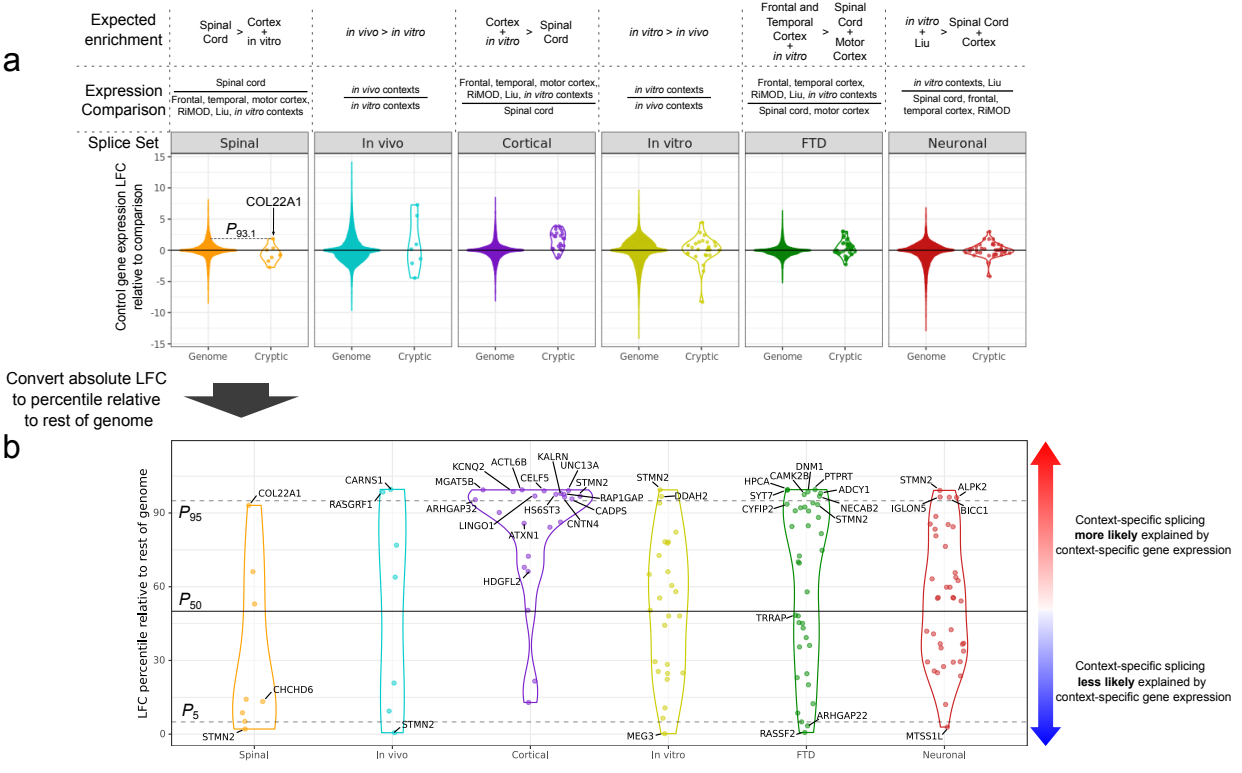

1. Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res 22:2008–2017. doi: 10.1101/gr.133744.111
2. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 20:163–172. doi: 10.1038/s41590-018-0276-y
3. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 57:289–300
4. Brettschneider J, Del Tredici K, Toledo JB, Robinson JL, Irwin DJ, Grossman M, Suh E, Van Deerlin VM, Wood EM, Baek Y, Kwong L, Lee EB, Elman L, McCluskey L, Fang L, Feldengut S, Ludolph AC, Lee VM-Y, Braak H, Trojanowski JQ (2013) Stages of pTDP-43 pathology in amyotrophic lateral sclerosis. Ann Neurol 74:20–38. doi: 10.1002/ana.23937
5. Brown A-L, Wilkins OG, Keuss MJ, Hill SE, Zanovello M, Lee WC, Bampton A, Lee FCY, Masino L, Qi YA, Bryce-Smith S, Gatt A, Hallegger M, Fagegaltier D, Phatnani H, Phatnani H, Kwan J, Sareen D, Broach JR, Simmons Z, Arcila-Londono X, Lee EB, Van Deerlin VM, Shneider NA, Fraenkel E, Ostrow LW, Baas F, Zaitlen N, Berry JD, Malaspina A, Fratta P, Cox GA, Thompson LM, Finkbeiner S, Dardiotis E, Miller TM, Chandran S, Pal S, Hornstein E, MacGowan DJ, Heiman-Patterson T, Hammell MG, Patsopoulos NikolaosA, Butovsky O, Dubnau J, Nath A, Bowser R, Harms M, Aronica E, Poss M, Phillips-Cremins J, Crary J, Atassi N, Lange DJ, Adams DJ, Stefanis L, Gotkine M, Baloh RH, Babu S, Raj T, Paganoni S, Shalem O, Smith C, Zhang B, Harris B, Broce I, Drory V, Ravits J, McMillan C, Menon V, Wu L, Altschuler S, Lerner Y, Sattler R, Van Keuren-Jensen K, Rozenblatt-Rosen O, Lindblad-Toh K, Nicholson K, Gregersen P, Lee J-H, Kokos S, Muljo S, Newcombe J, Gustavsson EK, Seddighi S, Reyes JF, Coon SL, Ramos D, Schiavo G, Fisher EMC, Raj T, Secrier M, Lashley T, Ule J, Buratti E, Humphrey J, Ward ME, Fratta P, NYGC ALS Consortium (2022) TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. Nature 603:131–137. doi: 10.1038/s41586-022-04436-3
6. Chen Y, Lun ATL, Smyth GK (2016) From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline
7. Conlon EG, Fagegaltier D, Agius P, Davis-Porada J, Gregory J, Hubbard I, Kang K, Kim D, New York Genome Center ALS Consortium, Phatnani H, Shneider NA, Manley JL (2018) Unexpected similarities between C9ORF72 and sporadic forms of ALS/FTD suggest a common disease mechanism. eLife 7:e37754. doi: 10.7554/eLife.37754
8. Germain P-L, Lun A, Garcia Meixide C, Macnair W, Robinson MD (2022) Doublet identification in single-cell sequencing data using scDblFinder. F1000Research 10:979. doi: 10.12688/f1000research.73600.2
9. Gittings LM, Alsop EB, Antone J, Singer M, Whitsett TG, Sattler R, 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) 146:433–450. doi: 10.1007/s00401-023-02599-5
10. Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, Gentleman R (2016) Prediction and Quantification of Splice Events from RNA-Seq Data. PLoS ONE 11:e0156132. doi: 10.1371/journal.pone.0156132
11. Gregoire Pau, Jens Reeder HTSeqGenie. In: Bioconductor. http://bioconductor.org/packages/HTSeqGenie/. Accessed 8 Jan 2024
12. Haghverdi L, Lun ATL, Morgan MD, Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36:421–427. doi: 10.1038/nbt.4091

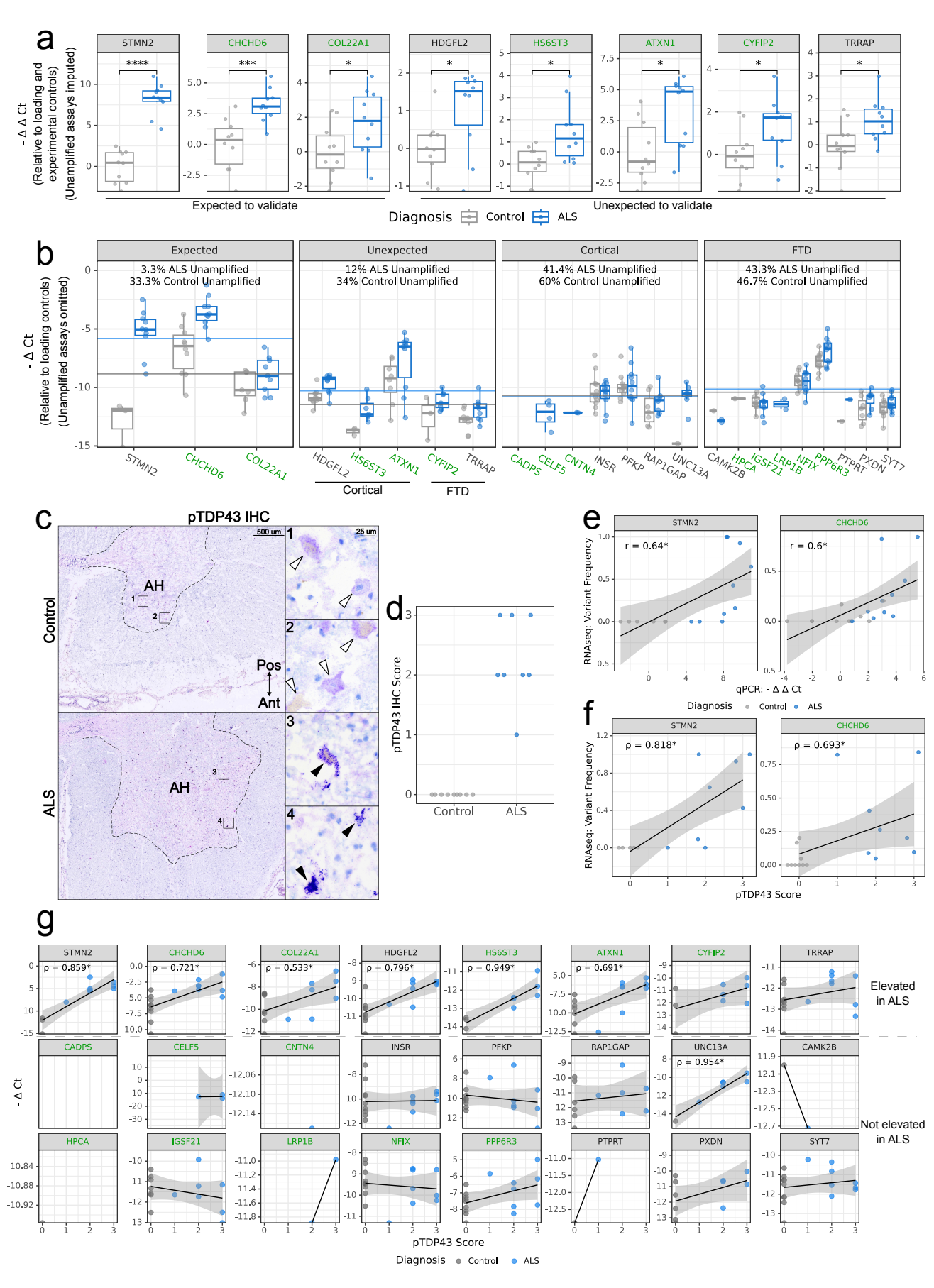
13. Irwin KE, Jasin P, Braunstein KE, Sinha I, Bowden KD, Moghekar A, Oh ES, Raitcheva D, Bartlett D, Berry JD, Traynor B, Ling JP, Wong PC (2023) A fluid biomarker reveals loss of TDP-43 splicing repression in pre-symptomatic ALS. BioRxiv Prepr Serv Biol 2023.01.23.525202. doi: 10.1101/2023.01.23.525202
14. Klim JR, Williams LA, Limone F, Guerra San Juan I, Davis-Dusenbery BN, Mordes DA, Burberry A, Steinbaugh MJ, Gamage KK, Kirchner R, Moccia R, Cassel SH, Chen K, Wainger BJ, Woolf CJ, Eggan K (2019) ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. Nat Neurosci 22:167–179. doi: 10.1038/s41593-018-0300-4
15. Li J, Jaiswal MK, Chien J-F, Kozlenkov A, Jung J, Zhou P, Gardashli M, Pregent LJ, Engelberg-Cook E, Dickson DW, Belzil VV, Mukamel EA, Dracheva S (2023) Divergent single cell transcriptome and epigenome alterations in ALS and FTD patients with C9orf72 mutation. Nat Commun 14:5714. doi: 10.1038/s41467-023-41033-y
16. Liao Y-Z, Ma J, Dou J-Z (2022) The Role of TDP-43 in Neurodegenerative Disease. Mol Neurobiol 59:4223–4241. doi: 10.1007/s12035-022-02847-x
17. Ling JP, Pletnikova O, Troncoso JC, Wong PC (2015) TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science 349:650–655. doi: 10.1126/science.aab0983
18. Liu EY, Russ J, Cali CP, Phan JM, Amlie-Wolf A, Lee EB (2019) Loss of Nuclear TDP-43 Is Associated with Decondensation of LINE Retrotransposons. Cell Rep 27:1409-1421.e6. doi: 10.1016/j.celrep.2019.04.003
19. Lun ATL, McCarthy DJ, Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor
20. Ma XR, Prudencio M, Koike Y, Vatsavayai SC, Kim G, Harbinski F, Briner A, Rodriguez CM, Guo C, Akiyama T, Schmidt HB, Cummings BB, Wyatt DW, Kurylo K, Miller G, Mekhoubad S, Sallee N, Mekonnen G, Ganser L, Rubien JD, Jansen-West K, Cook CN, Pickles S, Oskarsson B, Graff-Radford NR, Boeve BF, Knopman DS, Petersen RC, Dickson DW, Shorter J, Myong S, Green EM, Seeley WW, Petrucelli L, Gitler AD (2022) TDP-43 represses cryptic exon inclusion in the FTD–ALS gene UNC13A. Nature 603:124–130. doi: 10.1038/s41586-022-04424-7
21. McCarthy DJ, Campbell KR, Lun ATL, Wills QF (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 33:1179–1186. doi: 10.1093/bioinformatics/btw777
22. Mehta PR, Brown A-L, Ward ME, Fratta P (2023) The era of cryptic exons: implications for ALS-FTD. Mol Neurodegener 18:16. doi: 10.1186/s13024-023-00608-5
23. Melamed Z, López-Erauskin J, Baughn MW, Zhang O, Drenner K, Sun Y, Freyermuth F, McMahon MA, Beccari MS, Artates JW, Ohkubo T, Rodriguez M, Lin N, Wu D, Bennett CF, Rigo F, Da Cruz S, Ravits J, Lagier-Tourenne C, Cleveland DW (2019) Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. Nat Neurosci 22:180–190. doi: 10.1038/s41593-018-0293-z
24. Menden K, Francescatto M, Nyima T, Blauwendraat C, Dhingra A, Castillo-Lizardo M, Fernandes N, Kaurani L, Kronenberg-Versteeg D, Atasu B, Sadikoglou E, Borroni B, Rodriguez-Nieto S, Simon-Sanchez J, Fischer A, Craig DW, Neumann M, Bonn S, Rizzu P, Heutink P (2022) A multi-omics dataset for the analysis of Frontotemporal Dementia genetic subtypes. 2020.12.01.405894
25. Neumann M, Lee EB, Mackenzie IR (2021) FTLD-TDP pathological subtypes: clinical and mechanistic significance. Adv Exp Med Biol 1281:201–217. doi: 10.1007/978-3-030-51140-1_13
26. Ohm DT, Cousins KAQ, Xie SX, Peterson C, McMillan CT, Massimo L, Raskovsky K, Wolk DA, Van Deerlin VM, Elman L, Spindler M, Deik A, Trojanowski JQ, Lee EB, Grossman M, Irwin DJ (2022) Signature laminar distributions of pathology in frontotemporal lobar degeneration. Acta Neuropathol (Berl) 143:363–382. doi: 10.1007/s00401-021-02402-3
27. Pineda SS, Lee H, Fitzwalter BE, Mohammadi S, Pregent LJ, Gardashli ME, Mantero J, Engelberg-Cook E, DeJesus-Hernandez M, Blitterswijk M van, Pottier C, Rademakers R, Oskarsson B, Shah JS, Petersen RC, Graff-Radford NR, Boeve BF, Knopman DS, Josephs KA, DeTure M,

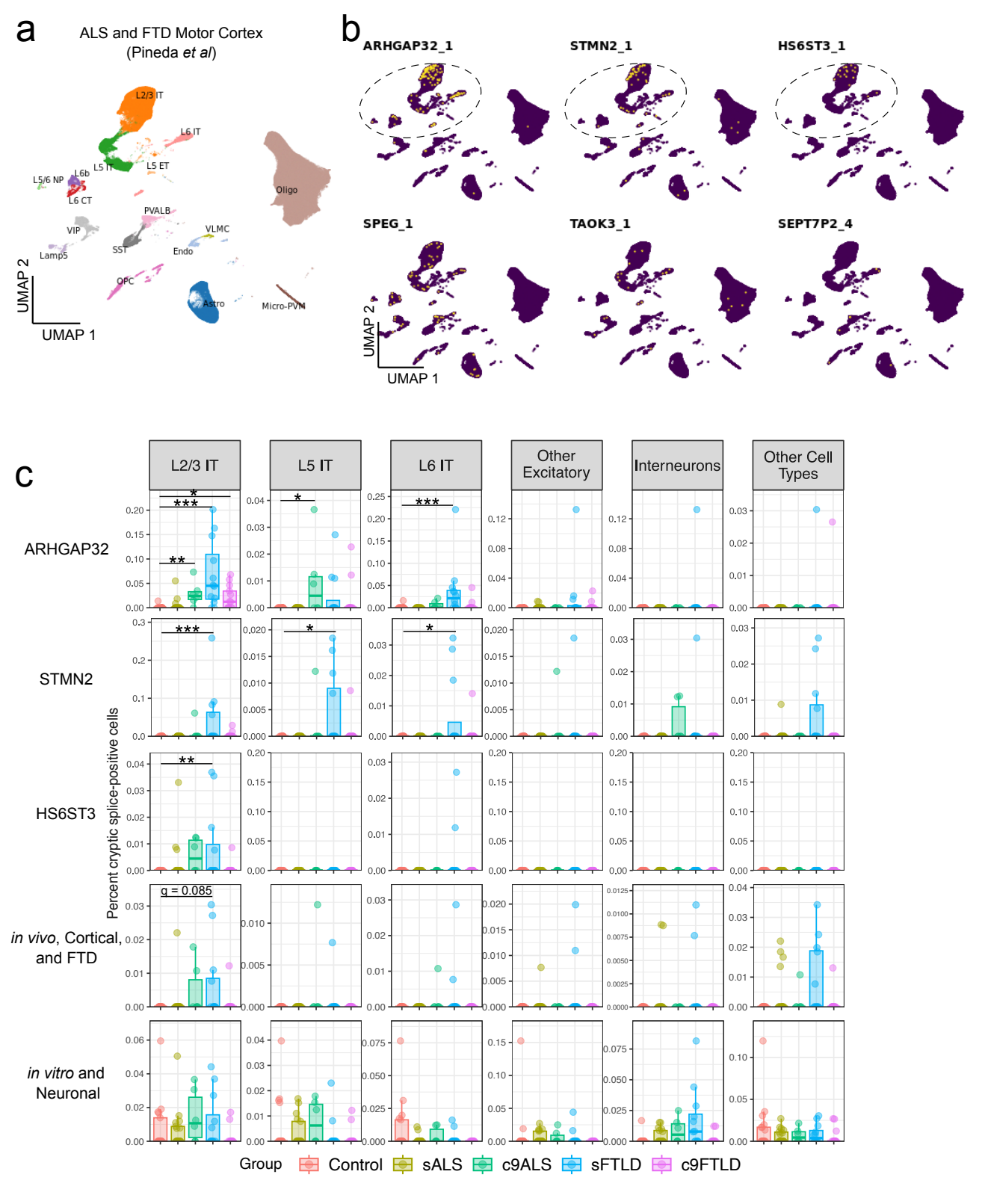
Murray ME, Dickson DW, Heiman M, Belzil VV, Kellis M (2021) Single-cell profiling of the human primary motor cortex in ALS and FTLD. 2021.07.07.451374
28. Prudencio M, Humphrey J, Pickles S, Brown A-L, Hill SE, Kachergus JM, Shi J, Heckman MG, Spiegel MR, Cook C, Song Y, Yue M, Daughrity LM, Carlomagno Y, Jansen-West K, de Castro CF, DeTure M, Koga S, Wang Y-C, Sivakumar P, Bodo C, Candalija A, Talbot K, Selvaraj BT, Burr K, Chandran S, Newcombe J, Lashley T, Hubbard I, Catalano D, Kim D, Propp N, Fennessey S, NYGC ALS Consortium, Fagegaltier D, Phatnani H, Secrier M, Fisher EM, Oskarsson B, van Blitterswijk M, Rademakers R, Graff-Radford NR, Boeve BF, Knopman DS, Petersen RC, Josephs KA, Thompson EA, Raj T, Ward M, Dickson DW, Gendron TF, Fratta P, Petrucelli L (2020) Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. J Clin Invest 130:6080–6092. doi: 10.1172/JCI139741
29. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. doi: 10.1093/bioinformatics/btp616
30. Seddighi S, Qi YA, Brown A-L, Wilkins OG, Bereda C, Belair C, Zhang Y, Prudencio M, Keuss MJ, Khandeshi A, Pickles S, Hill SE, Hawrot J, Ramos DM, Yuan H, Roberts J, Sacramento EK, Shah SI, Nalls MA, Colon-Mercado J, Reyes JF, Ryan VH, Nelson MP, Cook C, Li Z, Screven L, Kwan JY, Shantaraman A, Ping L, Koike Y, Oskarsson B, Staff N, Duong DM, Ahmed A, Secrier M, Ule J, Jacobson S, Rohrer J, Malaspina A, Glass JD, Ori A, Seyfried NT, Maragkakis M, Petrucelli L, Fratta P, Ward ME (2023) Mis-spliced transcripts generate de novo proteins in TDP-43-related ALS/FTD. 2023.01.23.525149
31. Srinivasan K, Friedman BA, Larson JL, Lauffer BE, Goldstein LD, Appling LL, Borneo J, Poon C, Ho T, Cai F, Steiner P, van der Brug MP, Modrusan Z, Kaminker JS, Hansen DV (2016) Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. Nat Commun 7:11295. doi: 10.1038/ncomms11295
32. Tam OH, Rozhkov NV, Shaw R, Kim D, Hubbard I, Fennessey S, Propp N, NYGC ALS Consortium, Fagegaltier D, Harris BT, Ostrow LW, Phatnani H, Ravits J, Dubnau J, Gale Hammell M (2019) Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. Cell Rep 29:1164-1177.e5. doi: 10.1016/j.celrep.2019.09.066
33. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40:e115. doi: 10.1093/nar/gks596
34. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881. doi: 10.1093/bioinformatics/btq057
35. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ (2016) GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In: Mathé E, Davis S (eds) Statistical Genomics: Methods and Protocols. Springer, New York, NY, pp 283–334
36. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 8 Jan 2024


a
GEO/SRA
EGA
GSE121569 PRJEB42763
GSE124439 GSE116622
GSE153960 GSE126543
EGAD00001008014
Ingest public bulk RNAseq
QC, Alignment
n=1,778
Spinal Cord
Cortex
Cerebellum
Hippocampus
Cervical
Thoracic
Lumbar
Motor
Frontal
Temporal
RiMOD
Liu
E-hMNs
NGN2
SH-SY5Y
SK-N-DZ
Control
ALS
ALS/FTD
FTLD-TDP43
FTLD-Tau
Non-ALS/FTD
TDP43+ nuclei
TDP43- nuclei
Scrambled siRNA
TDP43-KD siRNA
Sample Size
b
Event 1
Event 2
Event 3
Variant 1 Junctions
Variant 1 + Variant 2 Junctions
= Variant Frequency
ALS vs Control
FTLD-TDP vs Control
TDP43-KD vs scrambled siRNA
Identify Splice Events (SGSeq)
Differential splicing testing (DEXseq)
c
Significant
Not Significant
Cryptic
Non-Cryptic
APP: Event 5, Variant 2/4
STMN2: Event 2, Variant 1/2
ADD1: Event 10, Variant 2/2
q=3.58e-10
q=1.53e-5
Variant Frequency
Group
Control
ALS
ALS/FTD
Non-ALS/FTD
d
ALS
FTLD-TDP
TDP43-KD
Cryptic Variants
e
Retained Intron
Alternative Splice Site
Novel Exon Skip
Cryptic Casette Exon
Cryptic Terminal Exon
Cryptic Initial Exon
n= 142
Cryptic Splices
Previously-reported
Novel


a
NYGC ALS Consortium
FTD Frontal Cortex
Sorted TDP43+/- nuclei
in vitro TDP43-KD
Cerebellum
Hippocampus
Cervical Spinal Cord
Thoracic Spinal Cord
Lumbar Spinal Cord
Motor Cortex
Frontal Cortex
Temporal Cortex
RiMOD
Liu
E-hMNs
NGN2 neurons
SH-SY5Y
SK-N-DZ
Dataset
Group
Spinal
In vivo
Cortical
In vitro
FTD
Neuronal
STMN2_1_1/2
CHCHD6_1_2/2
COL22A1_1_1/2
KCNQ2_5_1/2
HS6ST3_1_1/2
KALRN_9_2/2
KALRN_30_2/2
HDGFL2_2_2/2
PFKP_2_2/2
ATAT1_1_2/2
UNC13A_6_2/2
MYO18A_5_1/2
SLC24A3_1_2/2
MYO1C_1_1/2
DNM1_9_1/2
SYT7_6_2/2
CAMK2B_6_2/2
CAMK2B_1_1/2
TRRAP_2_2/2
RSF1_1_2/2
CYFIP2_1_2/2
Relative Z-varFreq
6
4
2
0
Group
Control
ALS
ALS/FTD
FTLD-TDP43
FTLD-Tau
Non-ALS/FTD
TDP-43 positive
TDP-43 negative
scramble
TDP43-KD
b
STMN2
CHCHD6 (Spinal)
HS6ST3 (Cortical)
Variant Frequency
***
Lumbar Spinal Cord
Frontal Cortex
Group
Control
ALS
FTLD-TDP43
FTLD-Tau
Mean Coverage
Control
ALS
FTLD-TDP
FTLD-Tau
Lumbar Spinal Cord
Frontal Cortex
79610000
79630000
126957000
126957200
126957400
96090000
96100000
96110000
Ensembl Transcripts
Canonical
Cryptic
Chr8
Chr3
Chr13


a
Expected enrichment
Spinal Cord > Cortex + in vitro
in vivo > in vitro
Cortex + in vitro > Spinal Cord
in vitro > in vivo
Frontal and Temporal Cortex + in vitro > Spinal Cord + Motor Cortex
in vitro + Liu > Spinal Cord + Cortex
Expression Comparison
Spinal cord / Frontal, temporal, motor cortex, RiMOD, Liu, in vitro contexts
in vivo contexts / in vitro contexts
Frontal, temporal, motor cortex, RiMOD, Liu, in vitro contexts / Spinal cord
in vitro contexts / in vivo contexts
Frontal, temporal cortex, RiMOD, Liu, in vitro contexts / Spinal cord, motor cortex
in vitro contexts, Liu / Spinal cord, frontal, temporal cortex, RiMOD
Splice Set
Spinal
In vivo
Cortical
In vitro
FTD
Neuronal
Control gene expression LFC relative to comparison
COL22A1
P93.1
Genome
Cryptic
Convert absolute LFC to percentile relative to rest of genome
b
LFC percentile relative to rest of genome
P95
P50
P5
COL22A1
CHCHD6
STMN2
CARNS1
RASGRF1
KCNQ2
ACTL6B
KALRN
UNC13A
MGAT5B
CELF5
STMN2
RAP1GAP
ARHGAP32
HS6ST3
CADPS
LINGO1
CNTN4
ATXN1
HDGFL2
STMN2
ODAH2
MEG3
HPCA
CAMK2B
DNM1
PTPRT
SYT7
ADCY1
CYFIP2
NECAB2
STMN2
TRRAP
ARHGAP22
RASSF2
STMN2
ALPK2
IGLON5
BICC1
MTSS1L
Spinal
In vivo
Cortical
In vitro
FTD
Neuronal
Context-specific splicing more likely explained by context-specific gene expression
Context-specific splicing less likely explained by context-specific gene expression


a
STMN2
CHCHD6
COL22A1
HDGFL2
HS6ST3
ATXN1
CYFIP2
TRRAP
- Δ Δ Ct
(Relative to loading and experimental controls)
(Unamplified assays imputed)
Expected to validate
Unexpected to validate
Diagnosis
Control
ALS
b
Expected
Unexpected
Cortical
FTD
3.3% ALS Unamplified
33.3% Control Unamplified
12% ALS Unamplified
34% Control Unamplified
41.4% ALS Unamplified
60% Control Unamplified
43.3% ALS Unamplified
46.7% Control Unamplified
- Δ Ct
(Relative to loading controls)
(Unamplified assays omitted)
STMN2
CHCHD6
COL22A1
HDGFL2
HS6ST3
ATXN1
CYFIP2
TRRAP
CADPS
CELF5
CNTN4
INSR
PFKP
RAP1GAP
UNC13A
CAMK2B
HPCA
IGSF21
LRP1B
NFIX
PPP6R3
PTPRT
PXDN
SYT7
Cortical
FTD
c
pTDP43 IHC
500 um
25 um
Control
ALS
AH
Pos
Ant
d
pTDP43 IHC Score
Control
ALS
e
STMN2
CHCHD6
r = 0.64*
r = 0.6*
RNAseq: Variant Frequency
qPCR: - Δ Δ Ct
Diagnosis
Control
ALS
f
STMN2
CHCHD6
ρ = 0.818*
ρ = 0.693*
RNAseq: Variant Frequency
pTDP43 Score
g
STMN2
CHCHD6
COL22A1
HDGFL2
HS6ST3
ATXN1
CYFIP2
TRRAP
ρ = 0.859*
ρ = 0.721*
ρ = 0.533*
ρ = 0.796*
ρ = 0.949*
ρ = 0.691*
Elevated in ALS
CADPS
CELF5
CNTN4
INSR
PFKP
RAP1GAP
UNC13A
CAMK2B
ρ = 0.954*
HPCA
IGSF21
LRP1B
NFIX
PPP6R3
PTPRT
PXDN
SYT7
Not elevated in ALS
- Δ Ct
pTDP43 Score
Diagnosis
Control
ALS


a
ALS and FTD Motor Cortex
(Pineda et al)
L2/3 IT
L6 IT
L5 IT
L5 ET
L5/6 NP
L6b
L6 CT
Oligo
PVALB
VIP
VLMC
SST
Endo
Lamp5
OPC
Astro
Micro-PVM
UMAP 2
UMAP 1
b
ARHGAP32_1
STMN2_1
HS6ST3_1
SPEG_1
TAOK3_1
SEPT7P2_4
UMAP 2
UMAP 1
c
L2/3 IT
L5 IT
L6 IT
Other Excitatory
Interneurons
Other Cell Types
ARHGAP32
STMN2
HS6ST3
in vivo, Cortical, and FTD
in vitro and Neuronal
Percent cryptic splice-positive cells
q = 0.085
Group
Control
sALS
c9ALS
sFTLD
c9FTLD