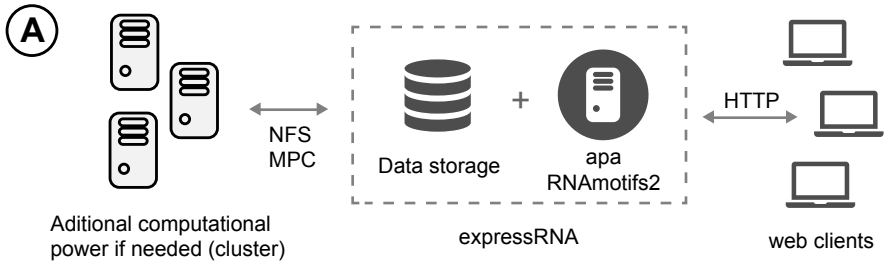
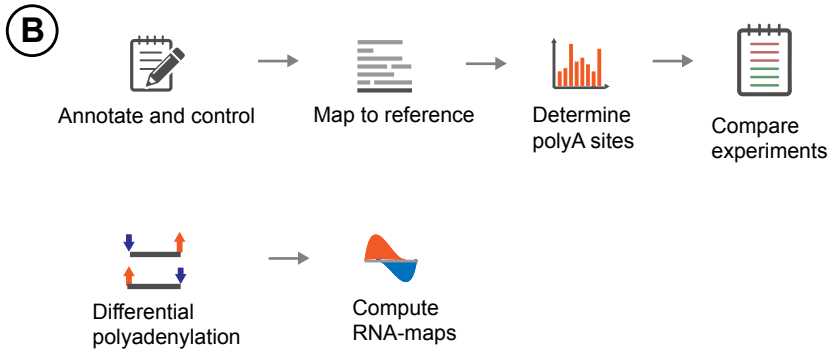
Cell Reports, Volume *19*

Supplemental Information

High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43

Gregor Rot, Zhen Wang, Ina Huppertz, Miha Modic, Tina Lenče, Martina Hallegger, Nejc Haberman, Tomaž Curk, Christian von Mering, and Jernej Ule





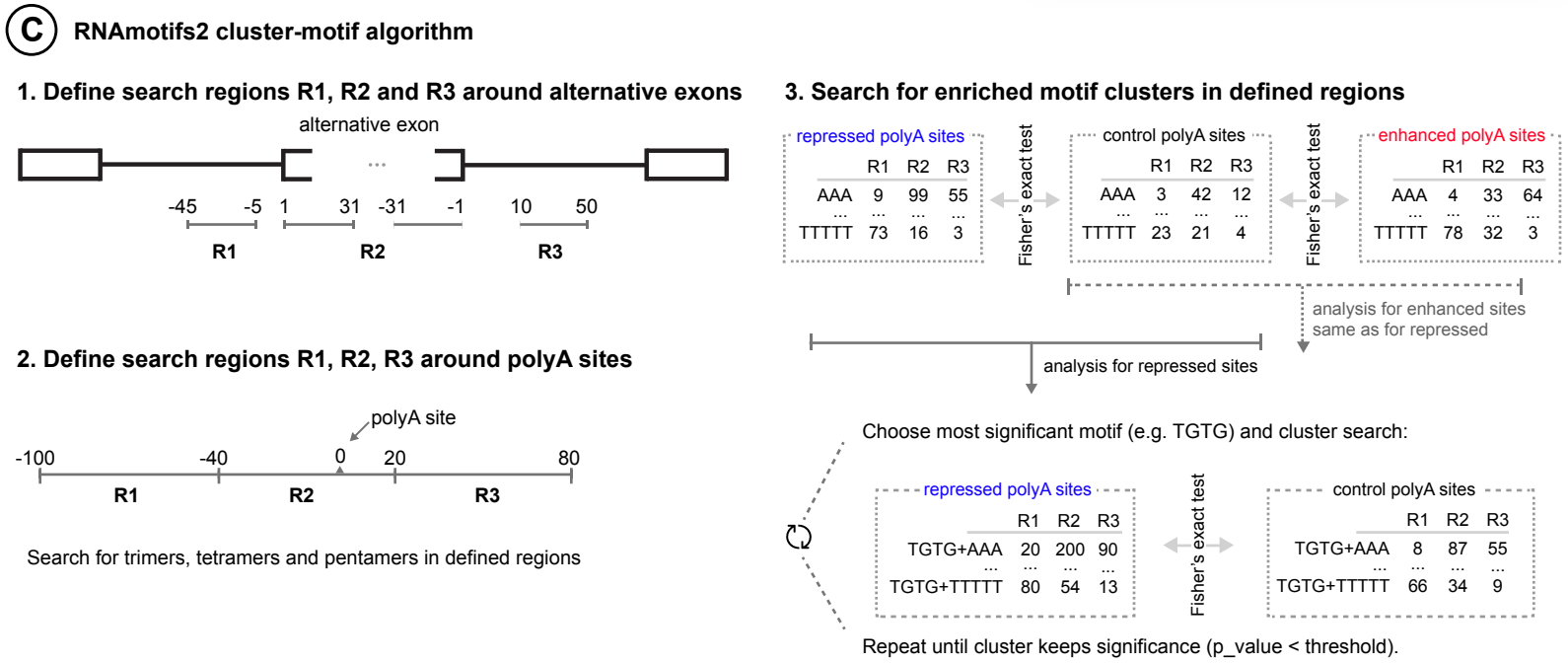


Figure S1. expressRNA research platform architecture with RNAmotifs2 schematic, Related to Figure 1

A. Platform processing architecture with mainframe server and data storage at the centre. **B.** 3'-end analysis workflow. **C.** RNAmotifs2: defined search regions for alternative splicing and alternative polyadenylation features, followed by the cluster search algorithm (for details see Supplemental Note). **D.** Screenshot of expressRNA.org web application, showing part of the results presented in this study.

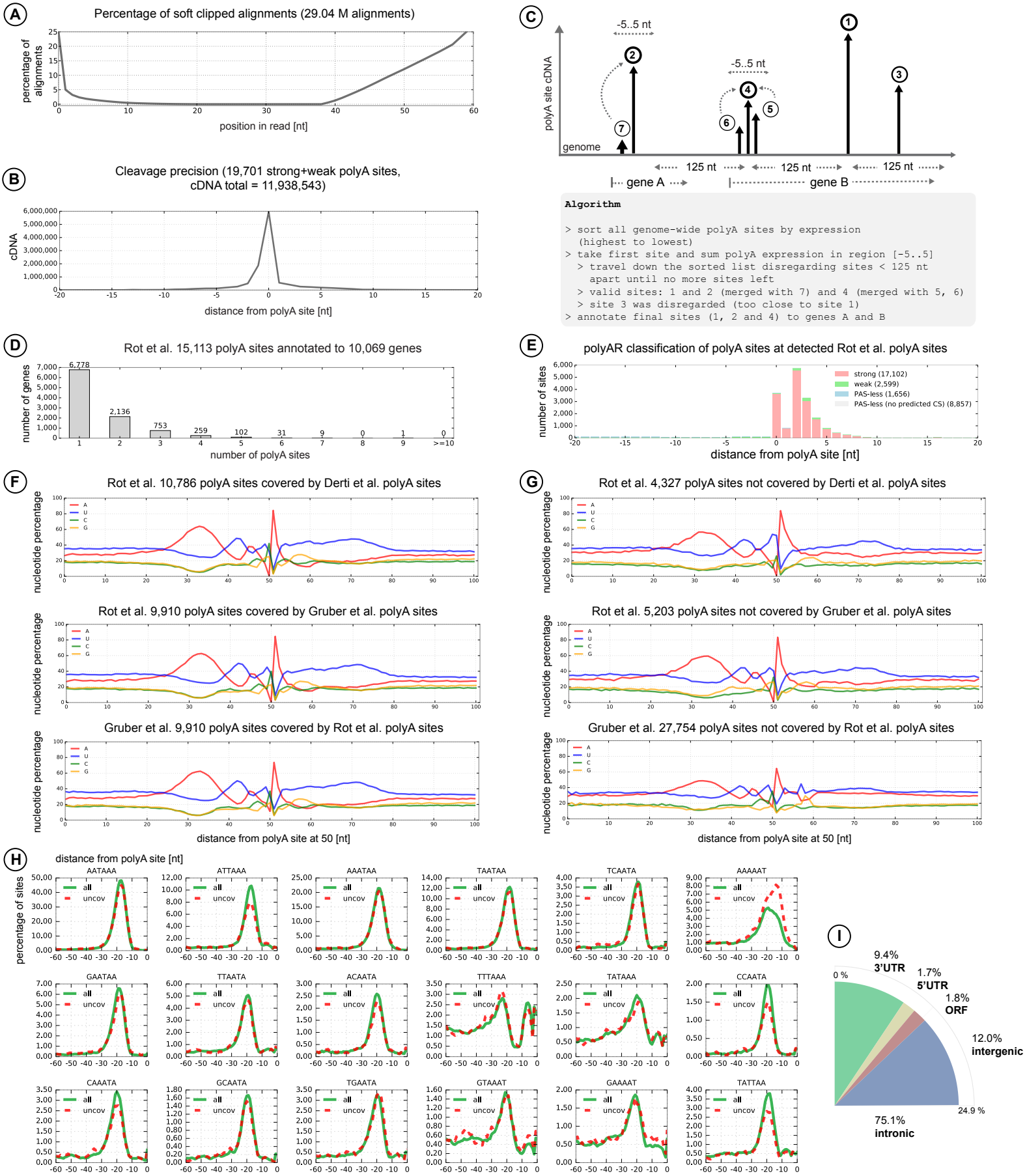
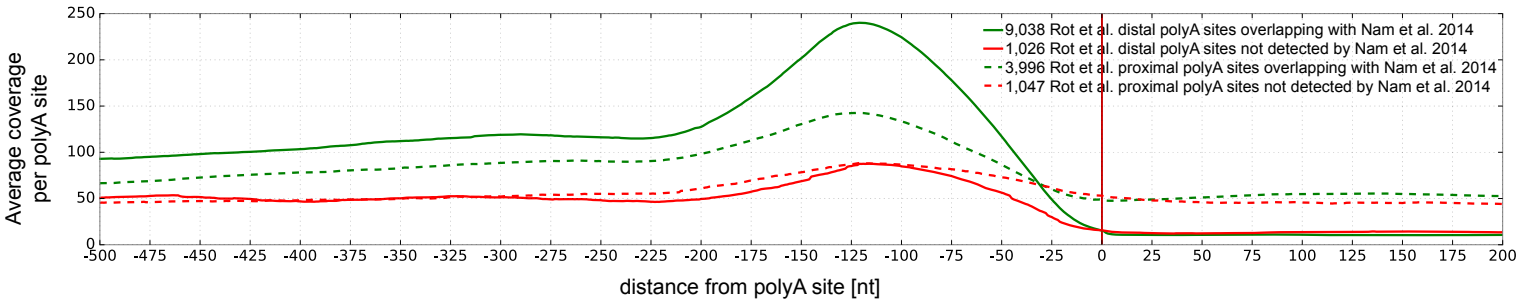


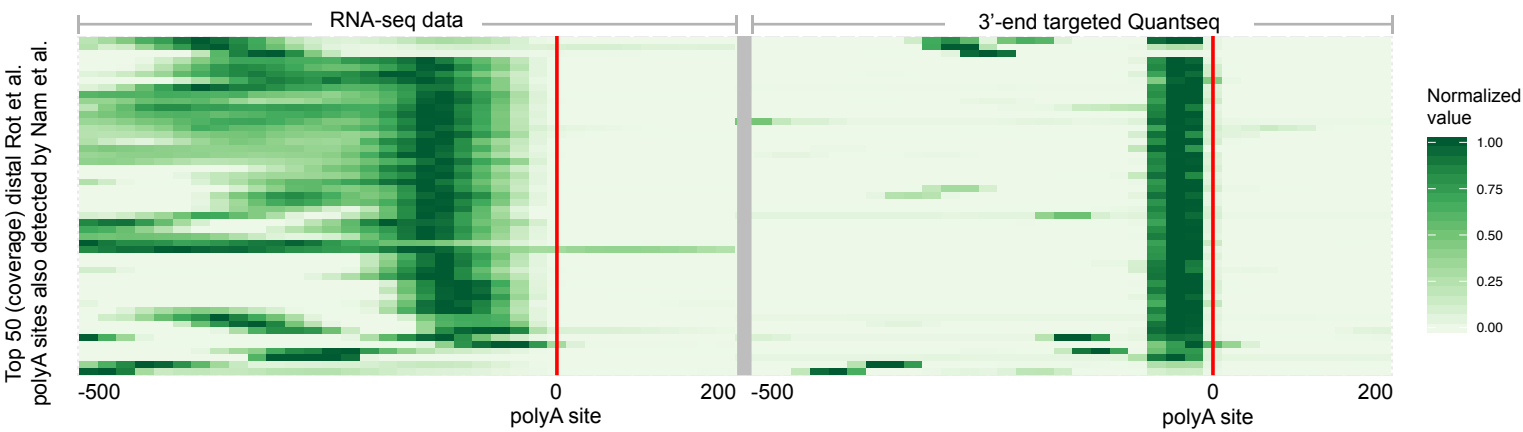
Figure S2. Polyadenylation and iCLIP analytics, Related to Figure 2

A. Soft clipping of alignments (5' due to sequencing into polyA tail, 3' due to imperfect primer annealing). **B.** Distribution of alignments around identified cleavage positions. Most alignments are within 5 nt of the polyA site. **C.** polyA database algorithm accounts for cleavage imprecision. **D.** Distribution of polyA sites and genes. **E.** polyAR software prediction of polyA sites at detected Rot et al. polyA sites (stacked). **F.** Nucleotide composition of Rot et al. polyA atlas covered by Derti et al. polyA sites and Gruber et al. polyA sites. Upstream A enrichment, downstream U enrichment and a peak at the cleavage site [50] all indicate the validity of the identified sites. **G.** Nucleotide composition around Rot et al. polyA sites not covered by Derti et al. polyA sites and Gruber et al. polyA sites. **H.** Presence of polyadenylation signals in region [-60, 0] around Rot et al. polyA sites (green: complete 15,113 Rot et al. polyA sites, red: only 5,203 Rot et al. polyA sites uncovered by Gruber et al). **I.** Genomic features distribution of 12,622,661 uniquely mapped TDP-43 iCLIP cDNAs.

J RNA-seq (11 HEK-293 experiments, 140M reads mapped) coverage of all proximal and distal Rot et al. polyA sites (green) and polyA sites not detected by Nam et al. 2014 (red).



K Top 50 distal Rot et al. polyA sites (genes with highest coverage) detected also by Nam et al. (2014). RNA-seq (11 HEK293 experiments) and Quantseq Reverse (12 experiments).



L Top 50 distal Rot et al. polyA sites (genes with highest coverage) not detected by Nam et al. (2014). RNA-seq (11 HEK293 experiments) and Quantseq Reverse (12 experiments).

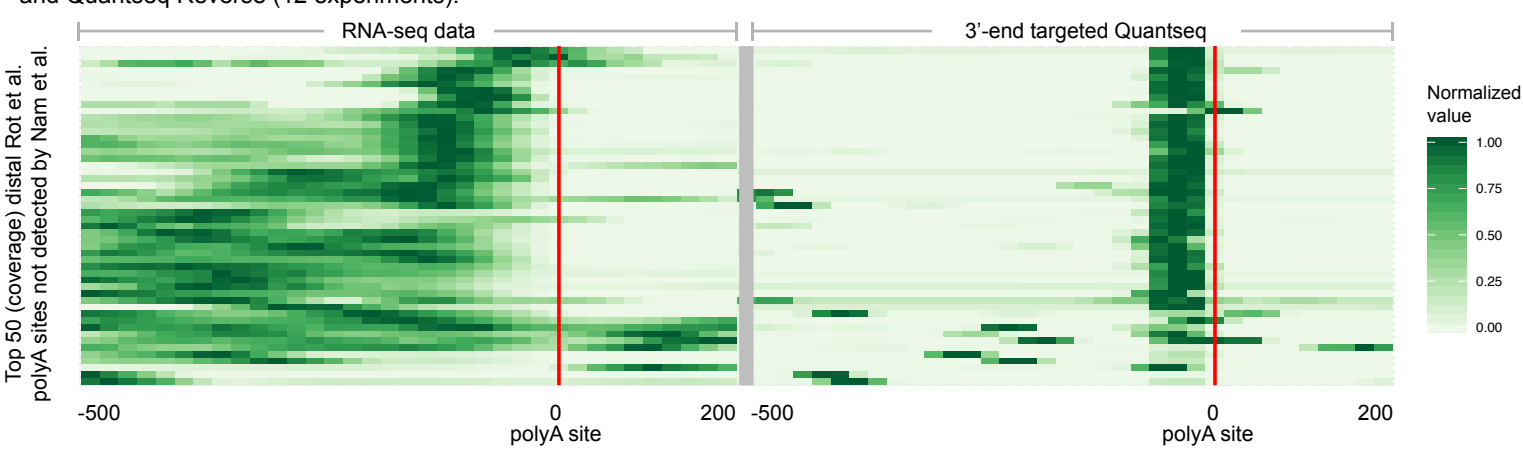


Figure S2 page 2. Validation of Rot et al. polyA sites with RNA-seq from GSE79680 and GSE82237), Related to Figure 2
J. Coverage of RNA-seq reads in the region [-500..200] around Rot et al. polyA sites. For distal sites the drop in coverage downstream of the polyA site at position 0 is indicative of the validity of detected sites. The coverage of polyA sites only detected by Nam et al. 2014 study is shown in red, however also these sites show decreased coverage in the downstream region (0..200). Proximal sites show less pronounced decrease in coverage as expected. **K.** Top 50 covered polyA sites heatmap with respective RNA-seq and Quantseq coverage (left and right panels). **L.** Same as K, however the top 50 covered polyA sites were selected only from polyA sites not detected by the Nam et al. 2014 study. The similar coverage distribution (compared to K.) validates our detected polyA sites not covered by Nam et al.

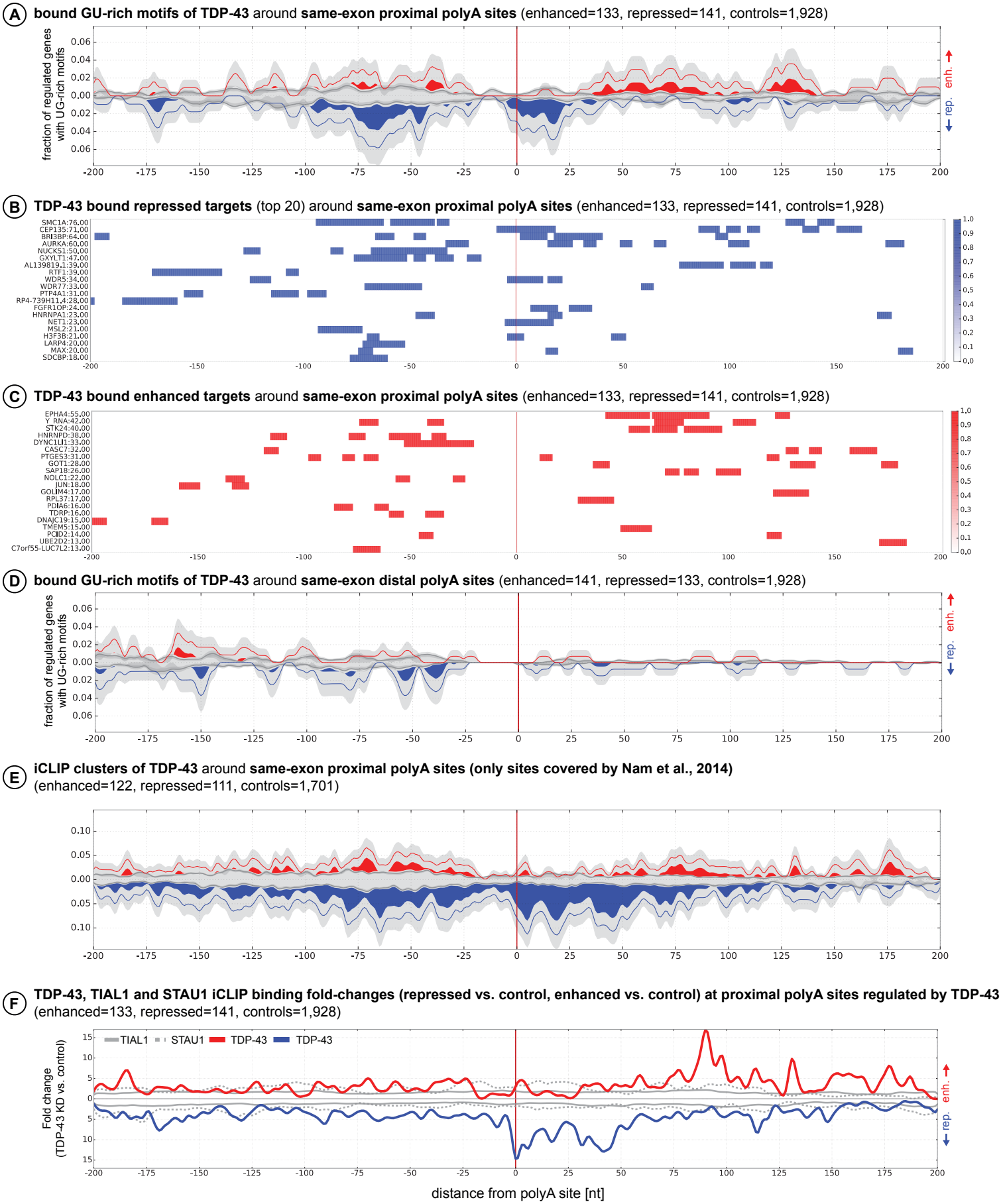


Figure S3. RNA-maps of TDP-43 around same-exon polyA sites with bound GU-rich motifs, Related to Figure 3

A. RNA map of UG-rich clusters at proximal site of TDP-43 bound regulated genes. TDP-43 binds in close proximity to repressed polyA sites (blue) and further downstream around enhanced polyA sites (red). **B.** Top 20 genes contributing to repression with corresponding bound UG-rich motif positions. **C.** Top genes contributing to enhancement with corresponding bound UG-rich motif positions. **D.** Distal site RNA map of TDP-43 binding, with less pronounced binding (Supplemental Table 6) compared to proximal site and controls (black lines). **E.** iCLIP RNA map (also see Figure 3A) is similar (robust) considering only polyA sites detected by Nam et al., 2014. **F.** TDP-43, TIAL1 and STAU1 iCLIP binding fold-changes (repressed/control_down, enhanced/control_up) around same-exon polyA sites regulated by TDP-43. The fold-changes coincide with UG-bound clusters (E) of TDP-43 regulated sites.

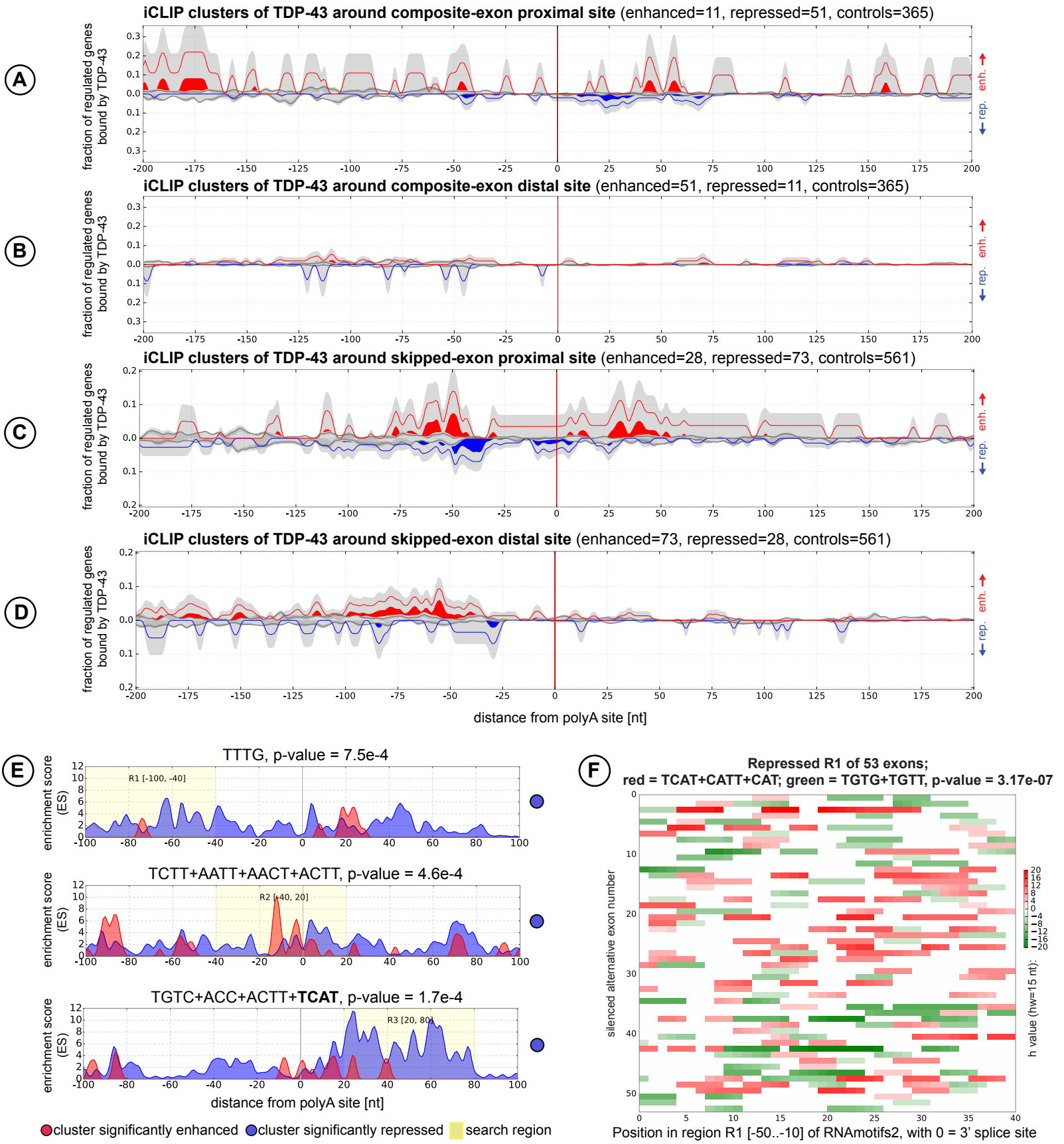


Figure S4. RNA maps of TDP-43 composite/skipped-exon; brain-specific motif clusters and UG/UC-rich motif cooccurence, Related to Figure 4

A. TDP-43 iCLIP at composite-exon proximal polyA sites. **B.** TDP-43 iCLIP at composite-exon distal polyA sites. **C.** TDP-43 iCLIP around skipped-exon proximal polyA sites. **D.** TDP-43 iCLIP at skipped-exon distal polyA sites. **E.** Most significant motif clusters (for each search region) regulating silencing of polyA sites in comparing brain and universal human reference (UHR). The TCAT motif is involved in alternative exon silencing. **F.** Coocurrence of UG-rich (green) and UC-rich (red) motif clusters at 53 exons in region R1 upstream of silenced alternative exons in comparing brain and heart (see also Figure 6). The h value is computed with a half-window of 15 nt and 53

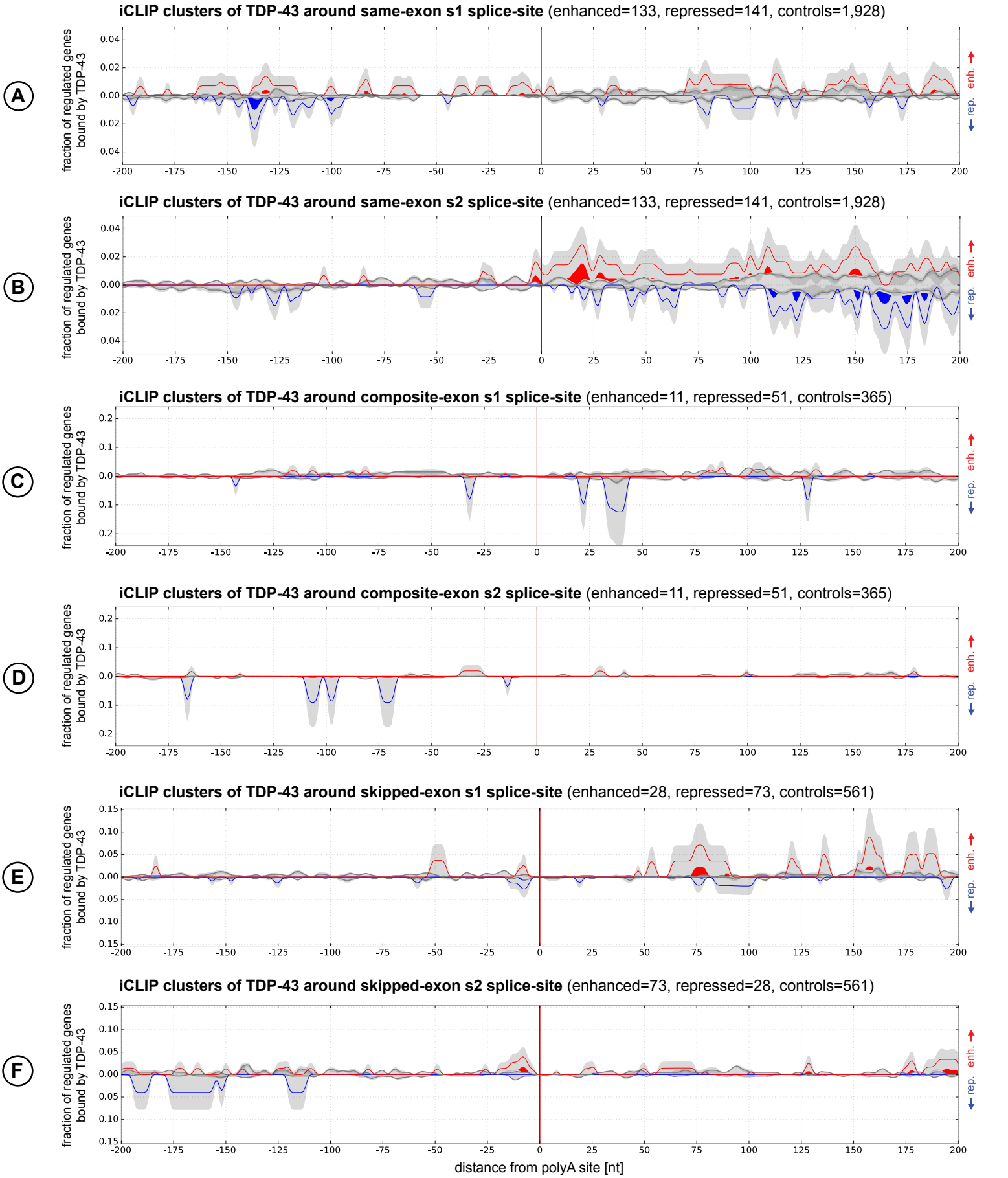


Figure S5. RNA maps of TDP-43 iCLIP around splice-sites, Related to Figure 3

A-F. The signal around splice-sites is very sparse and not enriched compared to controls (black lines). The number of bound regulated genes is lower (Supplemental Table 6) compared to bound regulated genes at polyA sites suggesting that TDP-43 is mostly not involved in the regulation of APA via splicing.

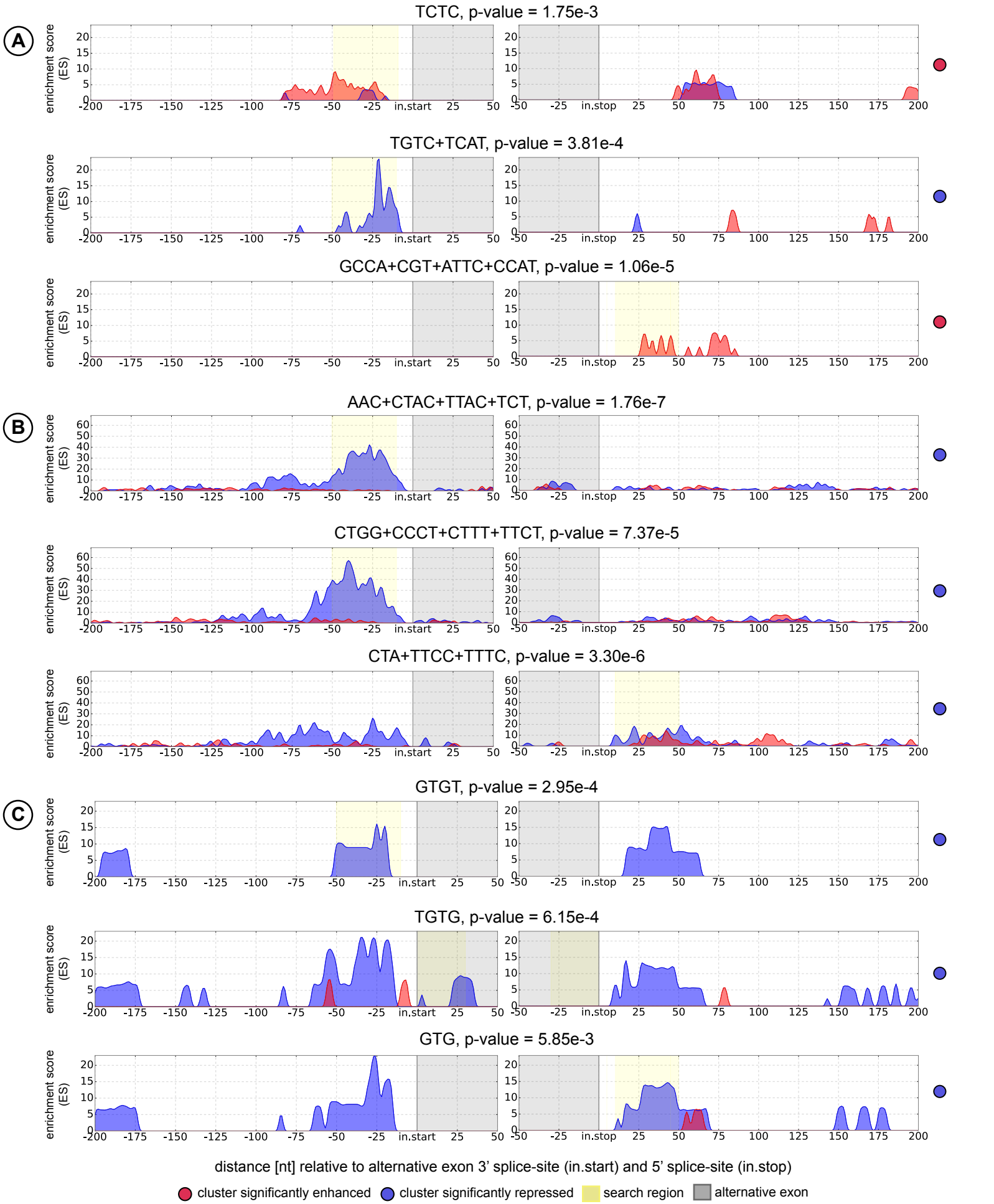


Figure S6. RNAmotifs2 cluster analysis on alternative splicing datasets, Related to Figure 6
A. NOVA protein clusters of regulated exons in NOVA$^{-/-}$ mouse brain neocortex splicing microarray. **B.** Analysis on PTBP regulated exons. **C.** Analysis on TDP-43 regulated exons. For details on splicing microarray datasets (A-C) see Cereda et al., 2014.

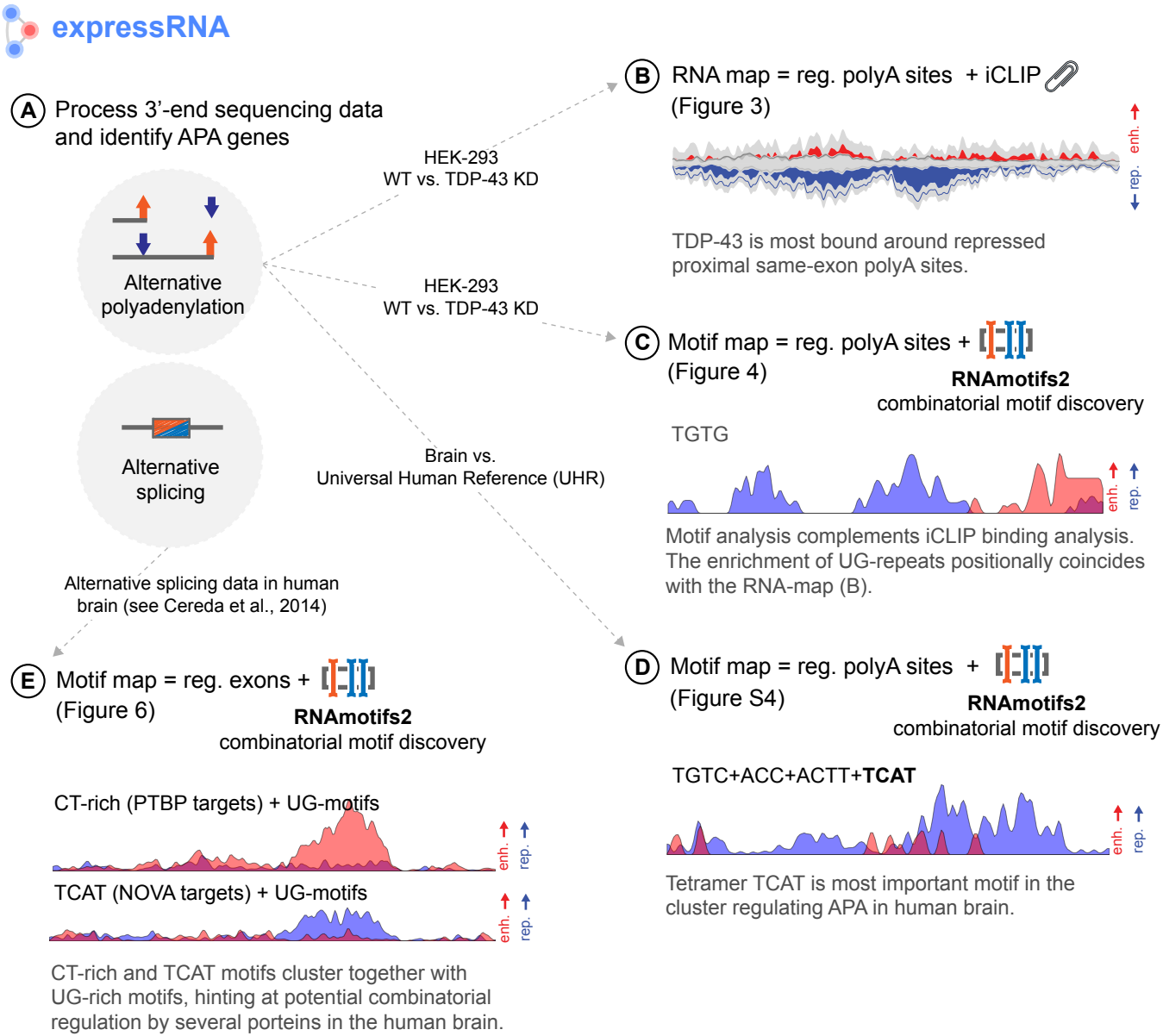


Figure S7. Overview of analysis and results, Related to Figure 1

A. Analysis of 3'-end sequence data with expressRNA. **B.** Integration with iCLIP data with information on APA genes discloses the role of TDP-43 in regulating polyA sites. **C.** Verification by the novel RNAmotifs2 analysis. **D.** The reconfirmed presence of TCAT at silenced polyA sites in the human brain. **E.** In depth cluster motif analysis alludes at the UG/UC-rich co-regulatory occurence, with reidentified TCAT in the human brain alternative splicing dataset.

RNAmotifs2 clustering algorithm

We extended the RNAmotifs software to account for clusters of short motifs. In addition to alternative splicing, we applied the software to search for motif clusters around regulated polyA sites. After retrieving the sequences of interest (R1, R2 and R3 around regulated features, either exons or polyA sites or some other features), we compute the search in several steps:

1. BASE motif search (motif runs all trimers, tetramers and pentamers)
- **a.** Make h_choosen such that closest to 4% (in 3-7% range) of all test features are detected
  - IF not possible to find h_choosen, skip motif
- **b.** Detect features using h_choosen
- **c.** Remove features with h>=14
- **d.** Fisher's exact test on detected features vs. all features
- **e.** Remember best motif as BASE_motif
  - Remember h_choosen as h_base
  - cluster = [BASE_motif] (single motif cluster)

2. IF Fisher(best motif) < 0.01, continue to step 3, else stop

3. Extend CLUSTER with N_motif (N_motif runs all trimers, tetramers and pentamers)
- **a.** Ignore features that were filtered out in previous steps
- **b.** Detect test features with [BASE_motif+N_motif] (both motifs must be present in the feature) using h = 0.5 * h_base; report detected features as SEARCH_features
- **c.** Make h_choosen such that closest to 4% (in range 3-7%) of test features and < 50% of SEARCH_features are covered
  - IF not possible to find h_choosen, skip to next N_motif
- **d.** Detect features from SEARCH_features with h_choosen
- **e.** Remove features from SEARCH_features with h>=14
- **f.** Fisher's exact test on detected features vs. all features

4. Stop IF Fisher(best cluster) > 0.001 or size of cluster >= 4
- **a.** Otherwise repeat step 3 and try to add another motif to the best cluster

5. Continue to **1**, disregarding already enriched motifs

The search is repeated for each region (R1, R2, R3) separatelly, and separatelly for silenced vs. control (1) and enhanced vs. control (2) comparisons. In total 3 * 2 = 6 searches are computed for each dataset. We consider only base motifs with p-value < 0.01 to start clustering, and clustering is halted when the p-value of the cluster > 0.001.

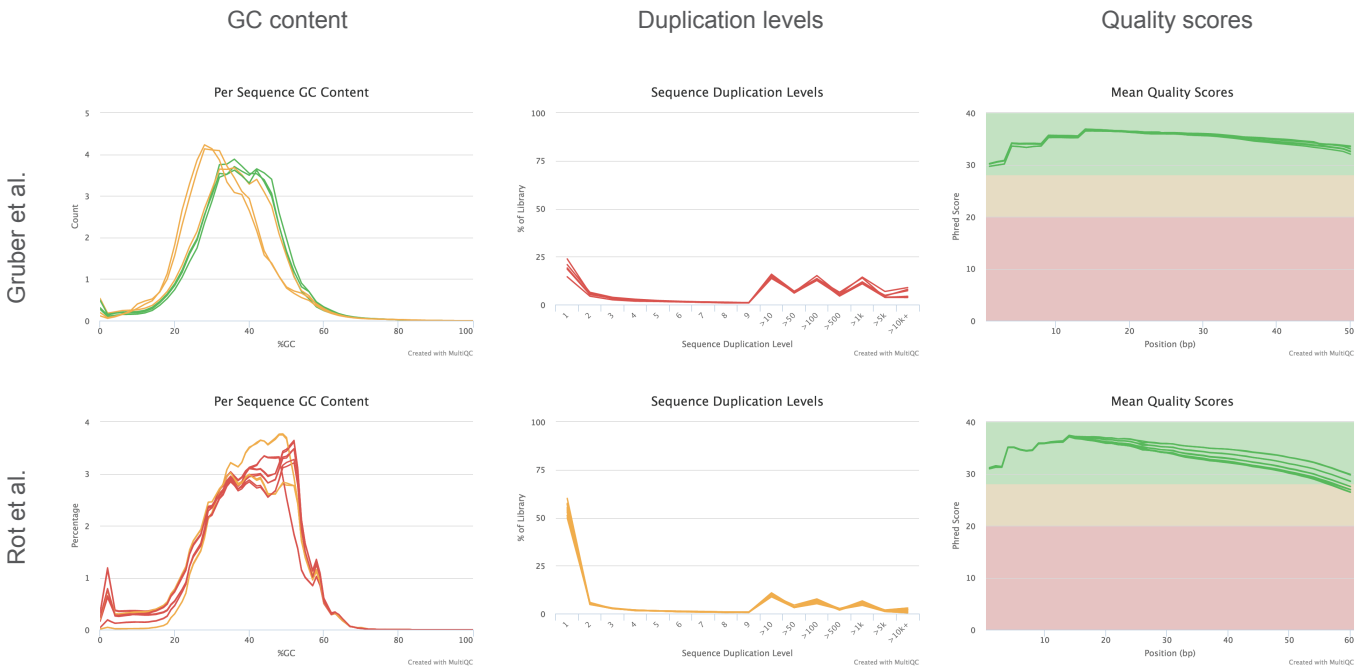
Comparing Lexogen QuantSeq reverse polyA atlas with Gruber et al., 2012 A-seq dataset

We applied the same analysis (expressRNA) as we describe in our manuscript to data from Gruber et al., RNA Biology, 2012. The Gruber et al. (2012) study quantified the choise of polyA sites in HEK-293 cells with an independent A-seq method. The analysis of the raw sequence data and consequent polyA database of our and the study by Gruber et al. (2012) confirms similarities on multiple levels: the mapability and the percentage of internal priming is very similar (55% mapability, 30% internally primed reads) and read quality and GC content is comparable between the two datasets (details in Supplemental Table 5).

Detecting significant hexamers (PAS) upstream of identified polyA sites

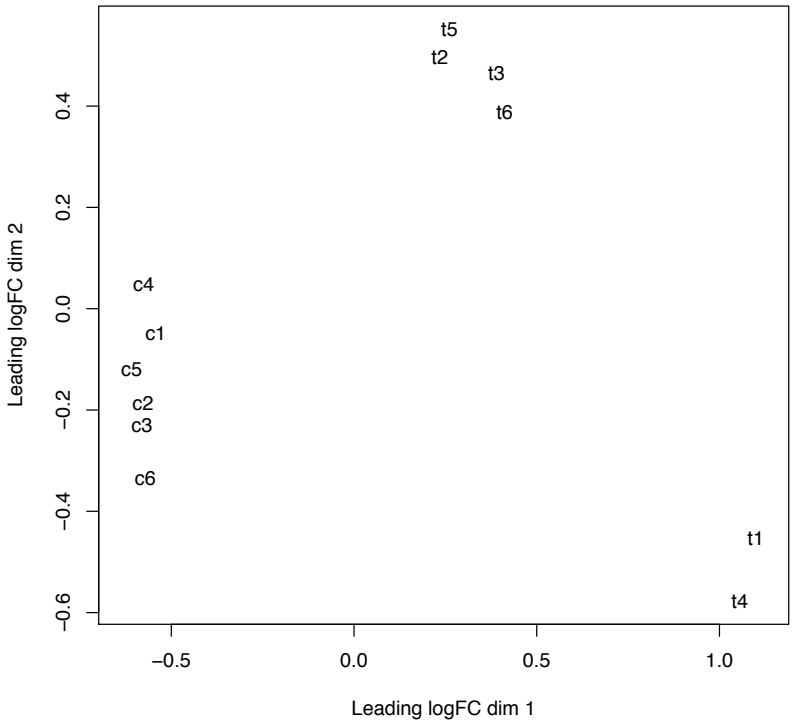
To identify the most relevant hexamers that are most likely to correspond to the polyA site (PAS) upstream of detected polyA sites in a de-novo fashion, we searched in regions (-30..-18) relative to polyA sites where the expected PAS signals are located. We compared the (-30..-18) signal to the signal in the control region (-60..-48), where we did not expect PAS elements. We then selected the top hexamer (Fisher's exact test) and plotted the presence of the hexamer in the (-30..-18) region (Figure S2H). For each hexamer search iteration, we removed the polyA sites with the hexamer present from future searches and repeated the search for the next hexamer.

Comparison to Gruber et al. 2012



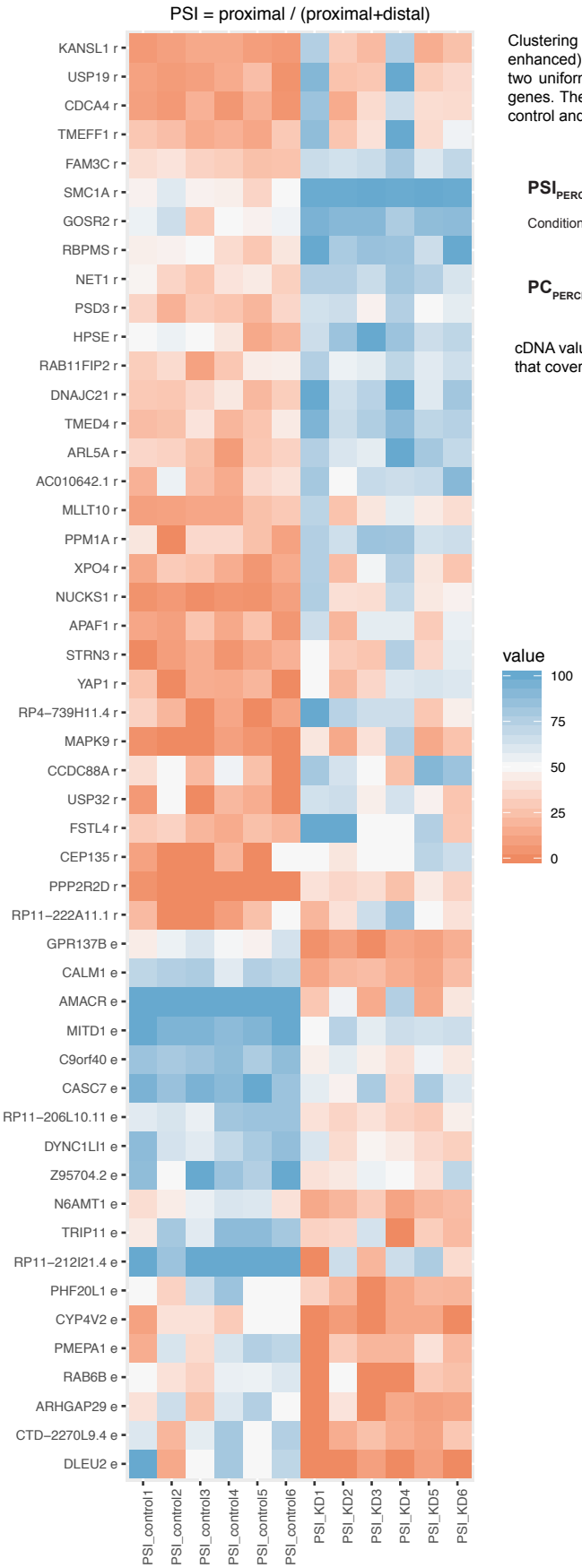
Sequence quality scores, GC content and sequence duplication levels (left to right) for the data from Gruber et al., 2012 and our study (top to bottom), analyzed with platform expressRNA.

MDS plot of 6 control and 6 KD replicates on the polyA site expression vectors



Separation of 6 control replicates (c1-c6) and 6 KD replicates (t1-t6). All control replicates cluster together and 4 of the KD replicates form a distinct cluster, with t4 and t1 further apart.

Per-replicate heatmap of PSI at proximal polyA sites



Clustering of 50 highest abs(PC) genes (r = repressed, e = enhanced). Clustering is performed with PSI values. Strikingly, two uniform clusters are formed, for repressed and enhanced genes. The variability in replicates show data is consistent and control and KD differences are quantifiable.

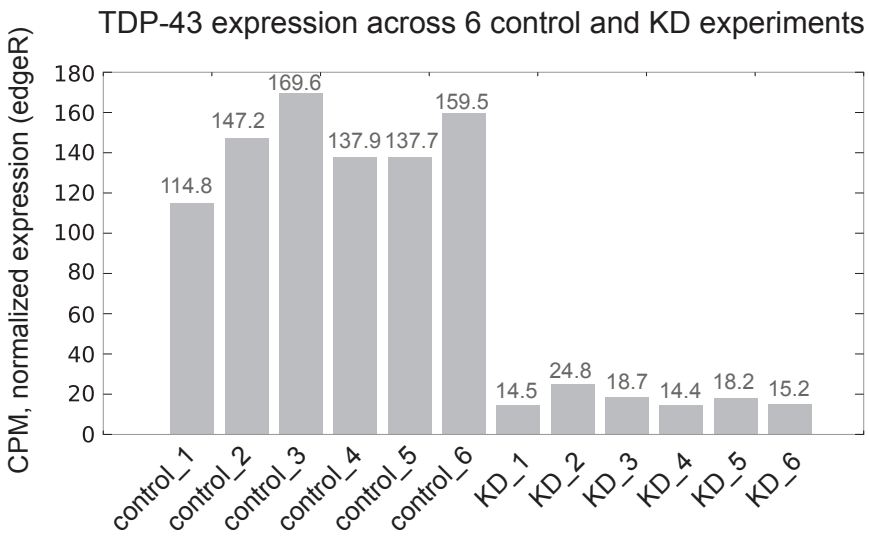
$$\mathbf{PSI}_{\text{PERCENT INCLUSION}}\,(\text{condition}) = \frac{\text{cDNA}_{\text{PROXIMAL}}}{\text{cDNA}_{\text{PROXIMAL+DISTAL}}}$$

Condition = control or KD

$$\mathbf{PC}_{\text{PERCENT CHANGE}} = \text{PSI(control) - PSI(test)}$$

cDNA values are the number of short-reads from each replicate that cover the polyA site (polyA site quantification).

Abundance of TARDBP mRNA in control and KD



Analysis of QuantSeq data confirms that the abundance of TARDBP mRNA is decreased by at least 80% in all replicates, with an 87% average decrease. We have validated that efficiency of TDP-43 knockdown was >80% with a western blot analysis (data not shown).

Occurence of top enriched hexamers (polyadenylation signals) around strong+weak (polyAR classified) and PAS-less Rot et al. polyA sites