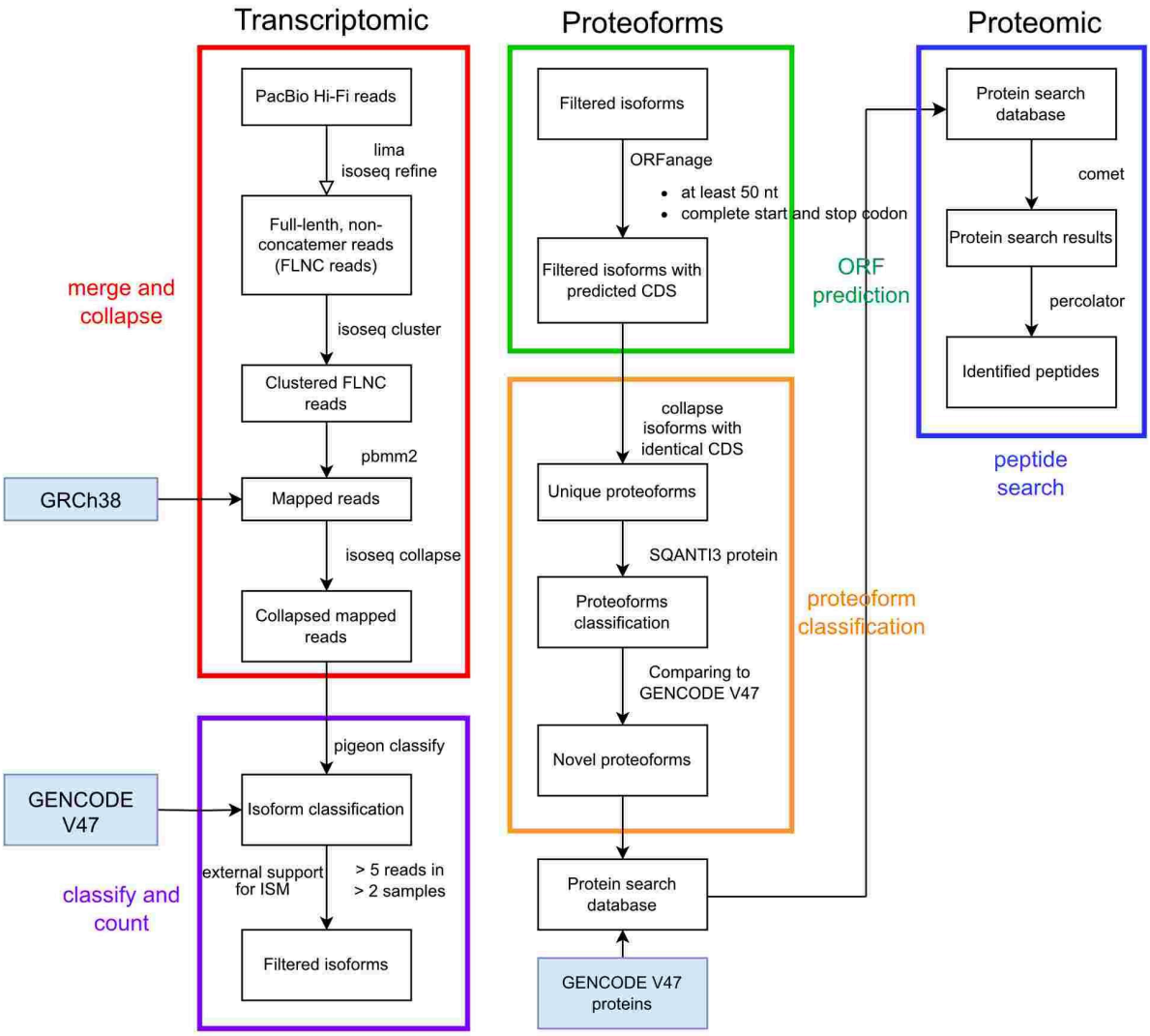
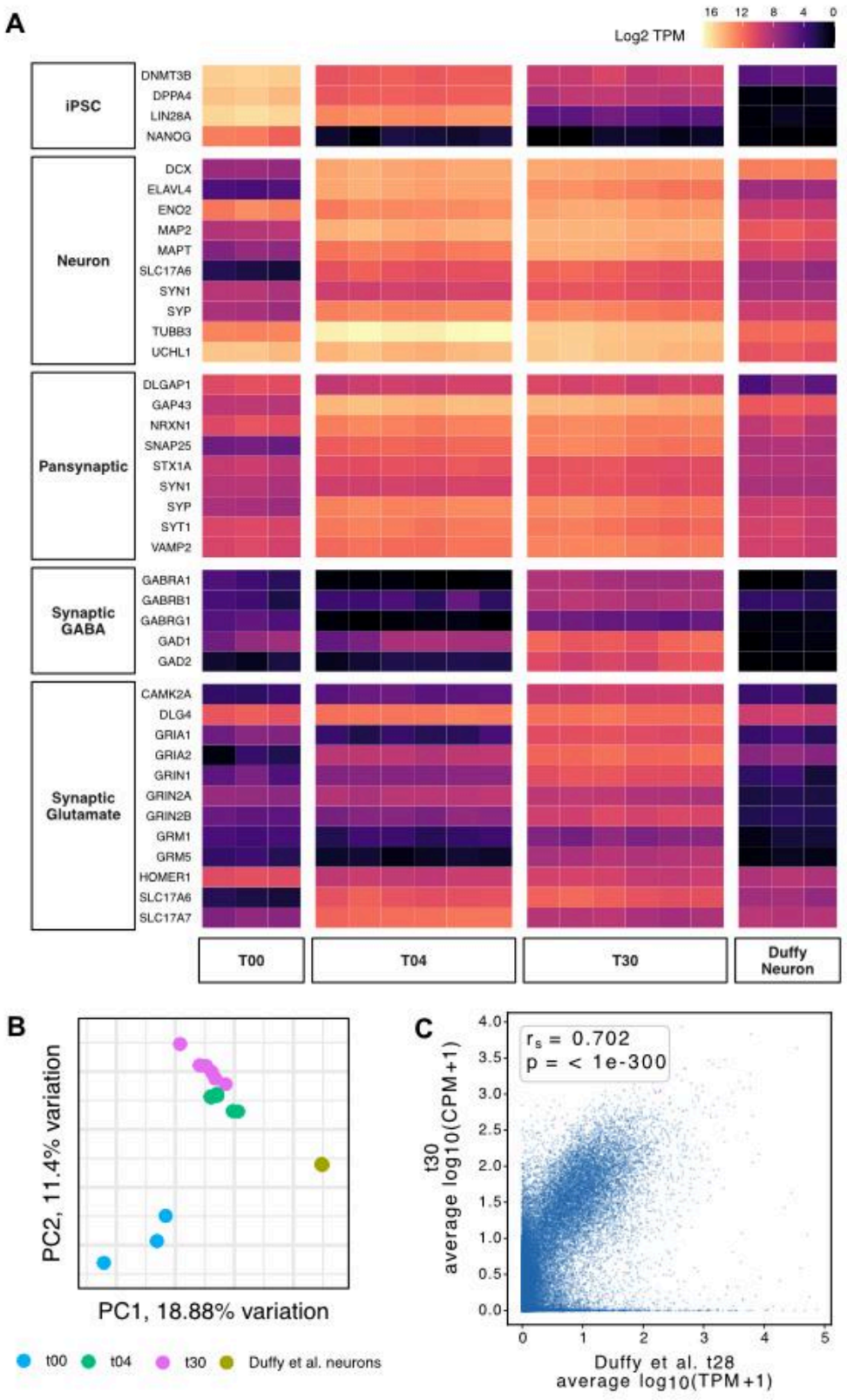
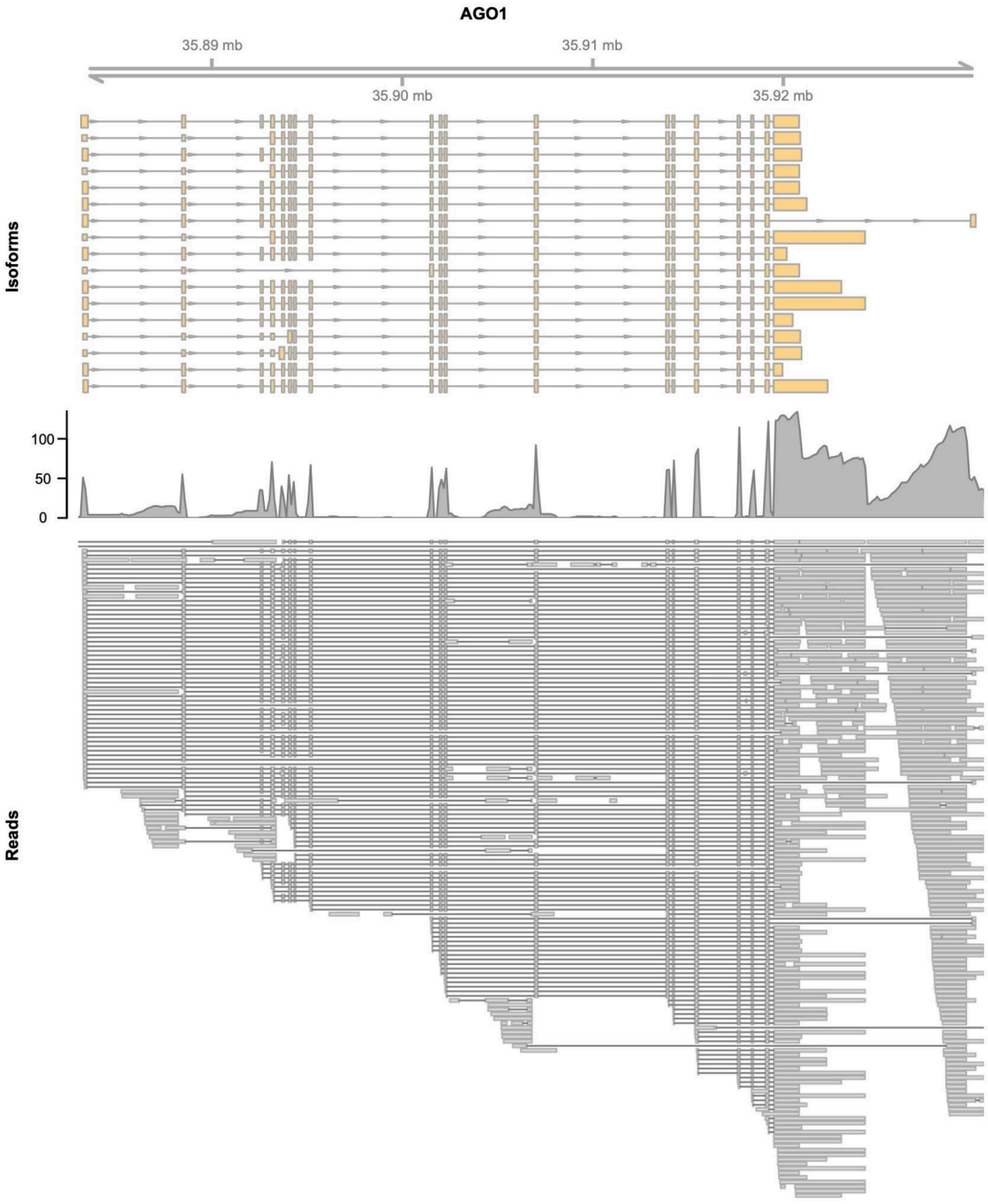
Long-read proteogenomic atlas of human neuronal differentiation reveals isoform diversity informing neurodevelopmental risk mechanisms



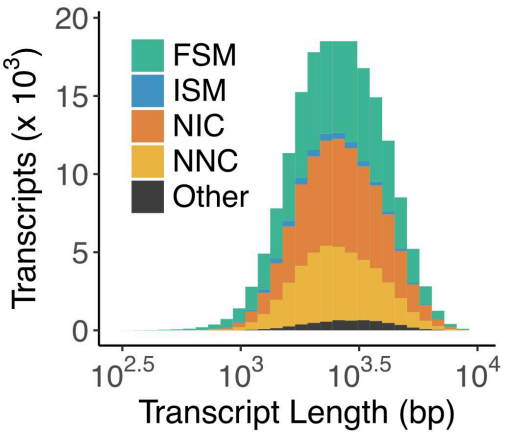
Supplementary Figure 1. An overview of the proteogenomic bioinformatics pipeline. This flowchart details the processing of raw long-read transcriptomic data (red box) and its classification against GENCODE V47 to define a high-confidence set of filtered isoforms (purple box). This transcriptome is then used to predict open reading frames (ORFs) (green box) and classify novel proteoforms (orange box). These novel proteoforms are combined with reference proteins to create a custom database used for the proteomic peptide search (blue box) to validate translation. Further details are provided in Methods.


A
Log2 TPM
16 12 8 4 0
iPSC
DNMT3B
DPPA4
LIN28A
NANOG
Neuron
DCX
ELAVL4
ENO2
MAP2
MAPT
SLC17A6
SYN1
SYP
TUBB3
UCHL1
Pansynaptic
DLGAP1
GAP43
NRXN1
SNAP25
STX1A
SYN1
SYP
SYT1
VAMP2
Synaptic GABA
GABRA1
GABRB1
GABRG1
GAD1
GAD2
Synaptic Glutamate
CAMK2A
DLG4
GRIA1
GRIA2
GRIN1
GRIN2A
GRIN2B
GRM1
GRM5
HOMER1
SLC17A6
SLC17A7
T00
T04
T30
Duffy Neuron
B
PC2, 11.4% variation
PC1, 18.88% variation
t00
t04
t30
Duffy et al. neurons
C
r_s = 0.702
p = < 1e-300
t30 average log10(CPM+1)
Duffy et al. t28 average log10(TPM+1)

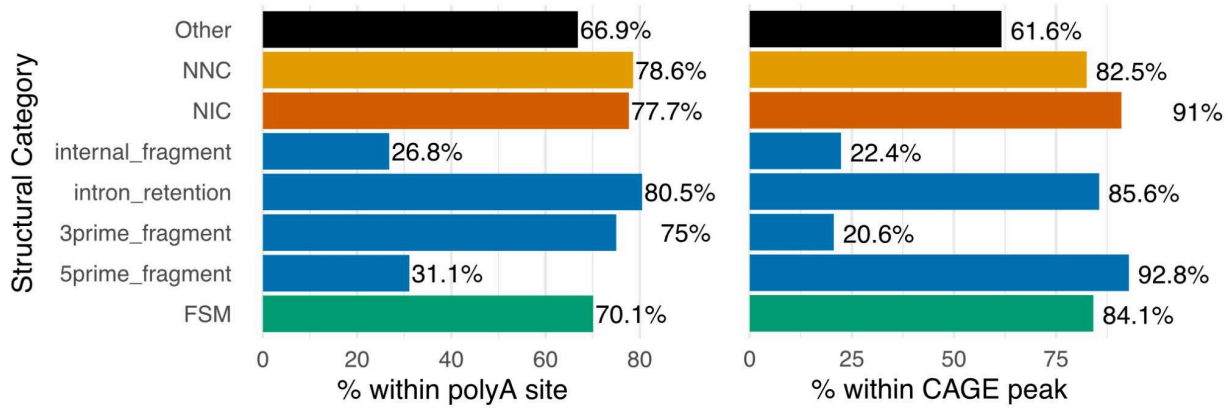
Supplementary Figure 2. Expression of canonical cell-type markers confirms cellular identity across the differentiation time course. (A) The heatmap shows gene expression (Log2 TPM) from this study's short-read RNAseq data at t00 (iPSC), t04 (intermediate), and t30 (neuron) time points. The final column displays expression from an external NGN2-based iPSC-derived neuron dataset for comparison[26]. (B) PCA was performed on log2-transformed TPM values across all samples in our study, and samples from the unstimulated group from Duffy et al. Each point represents an individual sample, colored by time point or study. The percentage of variance explained by each principal component is shown in parentheses on the respective axis. (C) Scatter plot comparing average gene expression between the unstimulated group from Duffy et al. and the t30 samples from our dataset. Each point represents one gene (n = 78,277). Both axes show $\log_{10}$(mean expression + 1). Spearman correlation coefficient (r_s) and associated p-value are indicated. Source data are provided as a Source Data file.



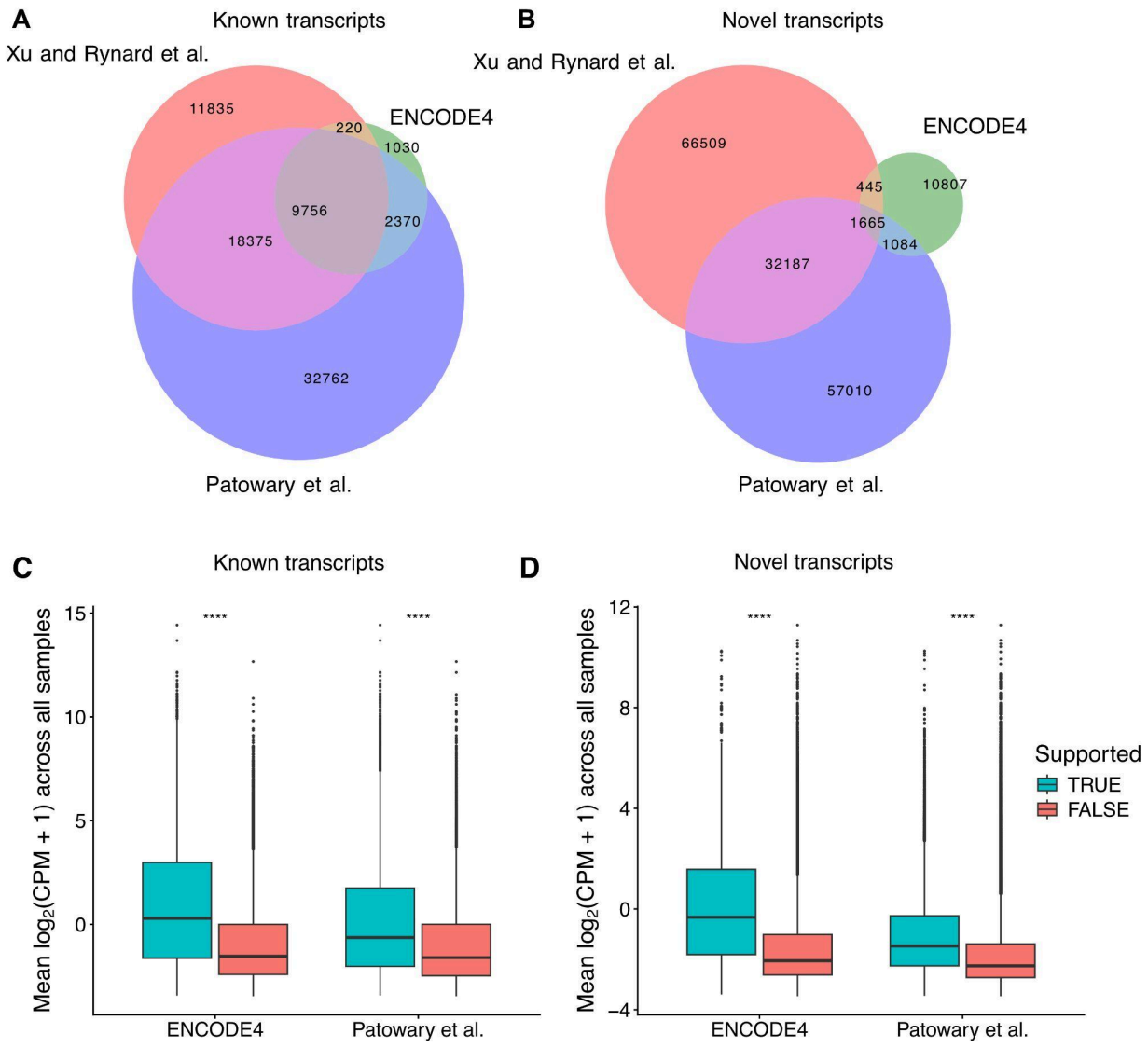
Supplementary Figure 3. *AGO1* isoforms and subsampled reads (10%) from this genomic region, pre-filtering are displayed. All *AGO1* isoforms are displayed.



Supplementary Figure 4. Distribution of isoform lengths from long-read RNAseq. Related to Fig. 1. This plot shows that isoforms of different structural categories have a similar distribution of transcript length that centres around 2600 bp. Source data are provided in Supplementary Data 1.

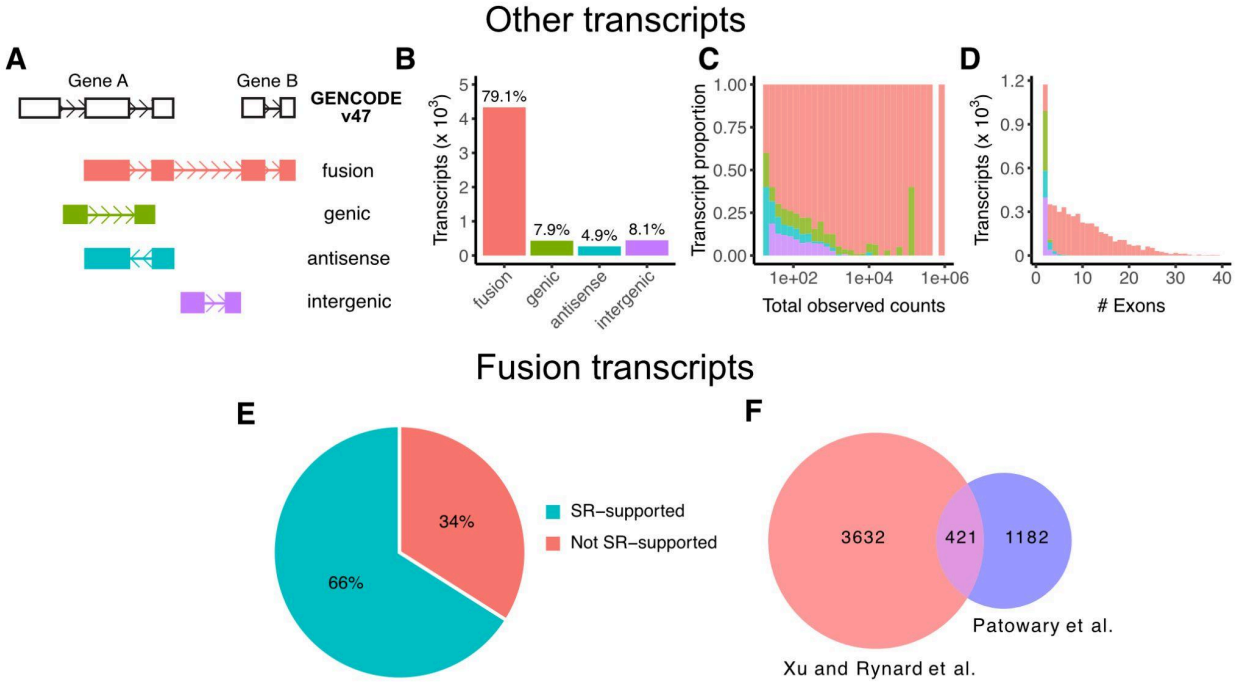


Supplementary Figure 5. Validation of isoform 3′ and 5′ ends using external datasets. Related to Fig. 1. Bar plots show the percentage of isoforms whose 3′ ends overlap a known polyA site (left) and whose 5′ ends overlap a CAGE peak (right). Structural categories match Fig. 1, with Incomplete Splice Matches (ISM) further stratified into four subcategories: internal fragments, intron retention, 3′ fragments (putative 5′ truncation), and 5′ fragments (putative 3′ truncation). To minimize technical artifacts, only 3′ fragments with valid 5′ support (CAGE peaks) and 5′ fragments with valid 3′ support (polyA sites) were retained for the final high-confidence catalog.

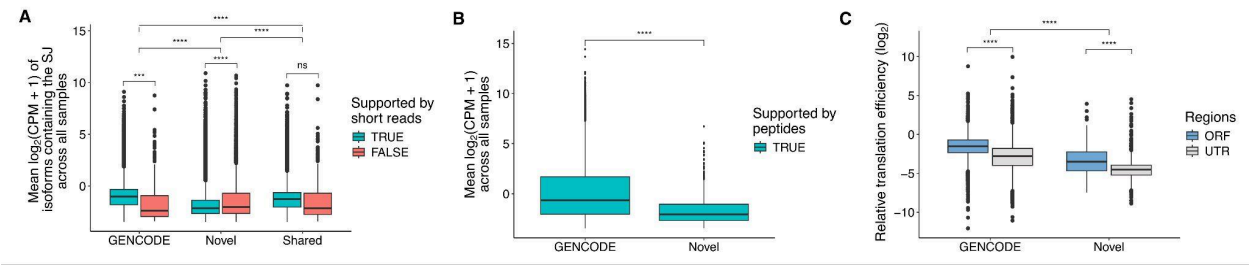


Supplementary Figure 6. Transcript abundance is a major contributor to replication of transcripts in external datasets. (A, B) Three-way Venn diagrams showing the overlap in transcripts detected in our dataset (Xu and Rynard et al.), and two external datasets, ENCODE4 (adult brain post-mortem tissues only) and Patowary et al (human embryonic brain tissues). Transcripts were compared by intron chain identity — two transcripts are considered equivalent if they share the same internal splice junctions, regardless of transcription start or end site variation. Only transcripts with more than 5 read counts in more than 2 samples were included. (A) Novel transcripts, defined as transcripts not matching any annotated isoform in the reference. (B) Known transcripts, defined as transcripts matching a reference isoform. Numbers indicate the count of transcripts unique to or shared between datasets. (C, D) Mean log2(CPM + 1) across all samples for transcripts in the current study, grouped by whether they are also present in ENCODE4 brain-expressed transcripts or in Patowary et al. transcripts. (C) Known transcripts. (D)

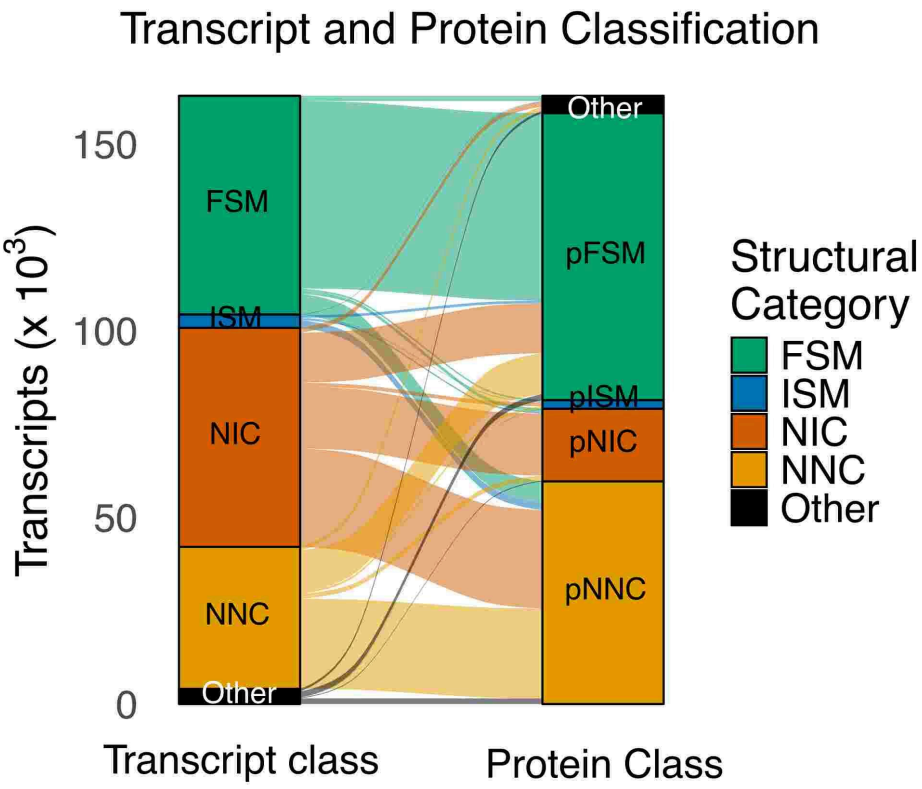
Novel transcripts. Significance of expression differences between supported and unsupported groups was assessed by two-sided Welch's t-test; ns, p > 0.05; *, p ≤ 0.05; **, p ≤ 0.01; ***, p ≤ 0.001; ****, p ≤ 0.0001. Boxes show the interquartile range with the median line; whiskers extend to 1.5× IQR. Source data are provided as a Source Data file.



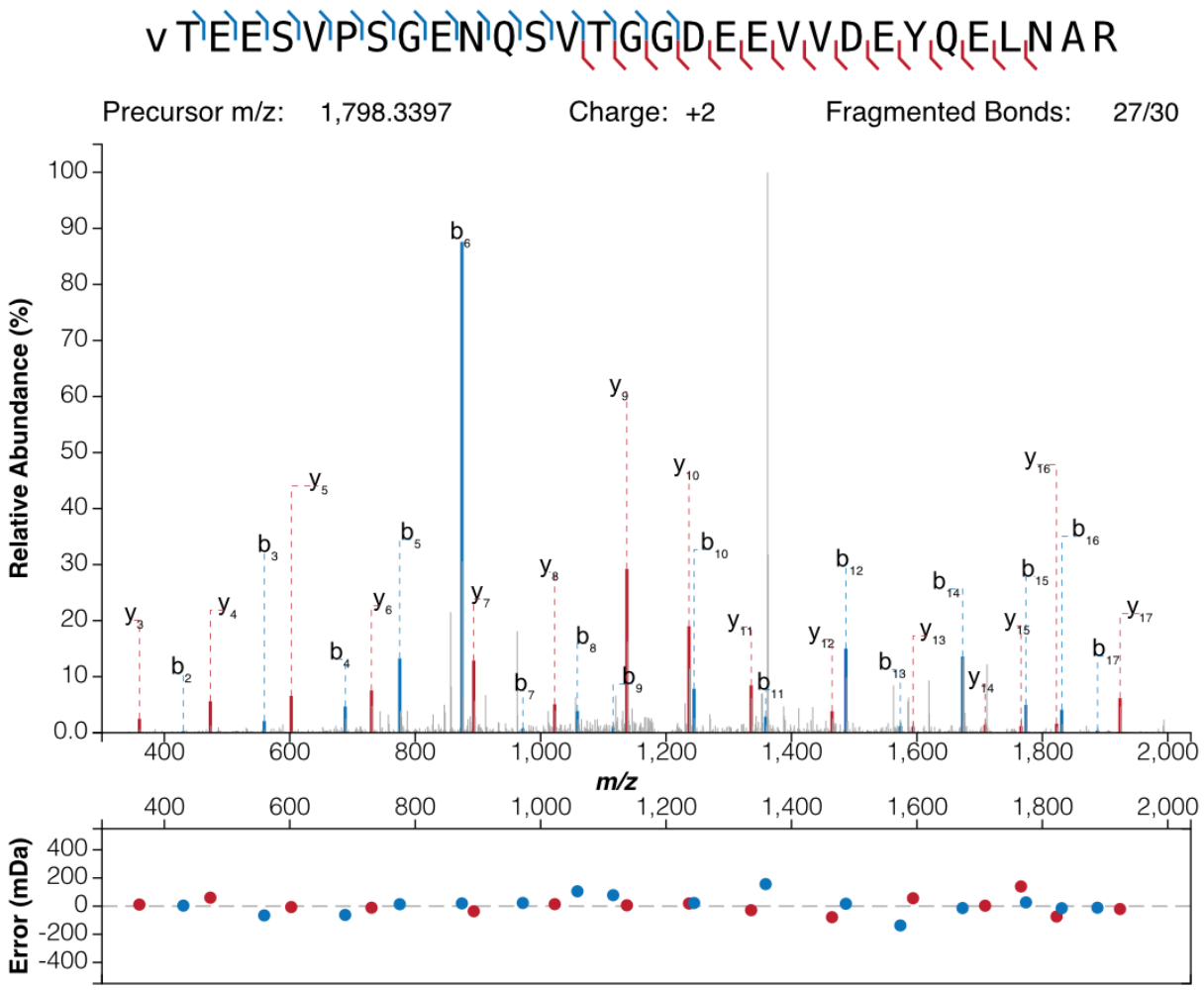
Supplementary Figure 7. Classification and characterization of fusion, genic, antisense, and intergenic transcripts. (A) Schematic illustrating the classification of transcripts relative to the GENCODE V47 reference. Categories are defined as 'fusion' (spanning two distinct gene loci), 'genic' (overlapping a known gene), 'antisense' (on the opposite strand of a known gene), and 'intergenic' (in a previously unannotated region). (B-D) Characterization of these transcript categories. These structural categories are used to illustrate the total count and proportion of all high-confidence transcripts (B), as well as the distributions of total read counts (C) and exon counts (D) per transcript. (E-F) Validation of fusion transcripts. (E) Pie chart showing the proportion of fusion transcripts (n = 4,332) whose splice junctions are fully supported by short-read RNA-seq data. 66.0% of fusion isoforms (2,860/4,332) have complete short-read splice junction support. (F) Venn diagram showing the overlap of fusion transcripts between our dataset and Patowary et al. dataset. 10.4% of fusion transcripts in our dataset are also found in Patowary et al. dataset. Source data are provided in Supplementary Data 1.



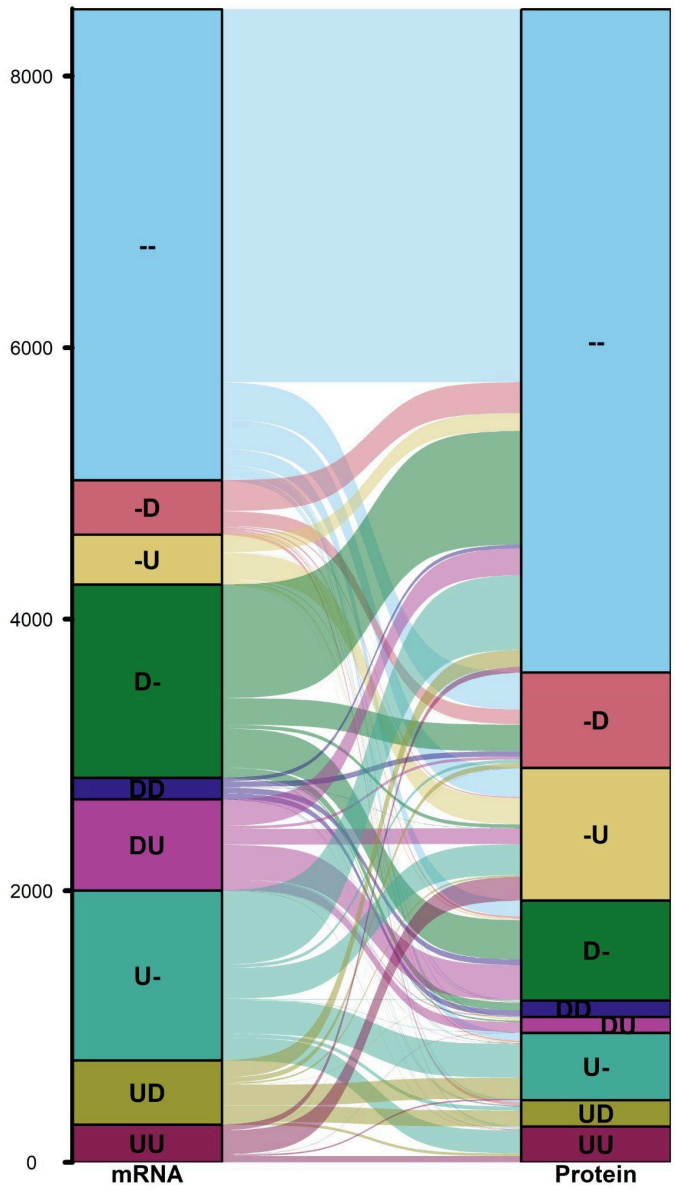
Supplementary Figure 8. Multi-omic validation of novel isoforms. (A) Mean expression of isoforms containing GENCODE, Novel, or Shared splice junctions, stratified by support from our matched short-read RNAseq. Validated junctions (teal) are associated with significantly higher isoform expression. (B) Mean expression of full-length isoforms supported by peptides from proteomics; where a peptide maps to multiple isoforms, expression is averaged across all isoforms containing peptide. Validated novel isoforms exhibit significantly lower expression than validated known isoforms. (C) Relative translational efficiency (TE), defined as the ratio of ribosome-protected fragments to input mRNA, calculated using external Ribo-seq data from human iPSC-derived neurons from Duffy et al., 2022. As a positive control, TE is significantly higher in known GENCODE coding sequences compared to 3′UTR. This pattern extends to the novel isoforms identified here, where predicted ORFs exhibit significantly greater TE compared to predicted 3′UTRs, providing independent validation for the ribosomal engagement and translation of these novel coding sequences. ***P < 0.001; ****P < 0.0001; ns = not significant. Source data are provided as a Source Data file.



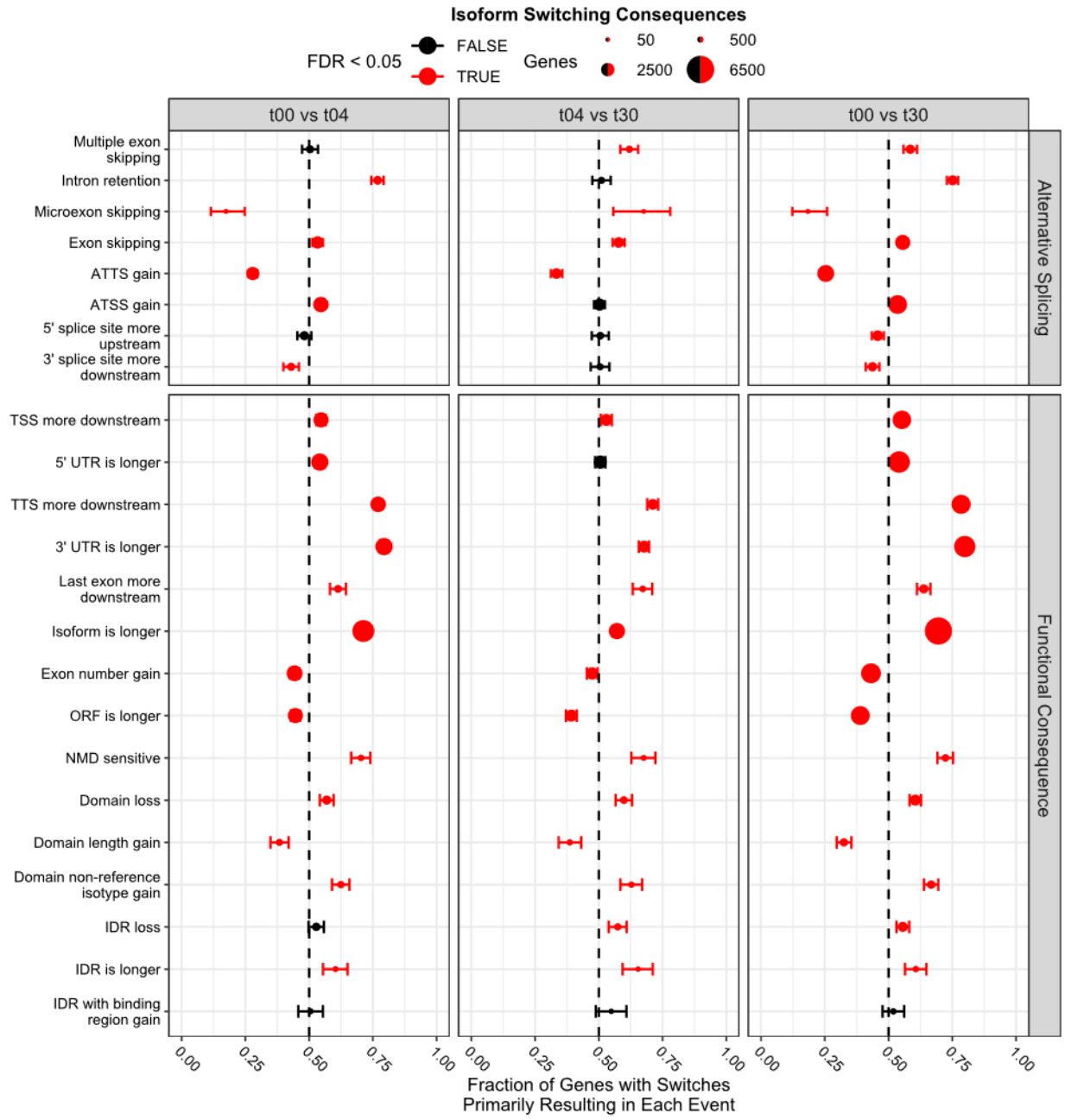
Supplementary Figure 9. Relationship between transcript and protein structural classifications. Sankey diagram illustrating the mapping between the structural classification of mRNA isoforms and their corresponding predicted protein isoforms, using the SQANTI protein framework. Source data are provided as a Source Data file.



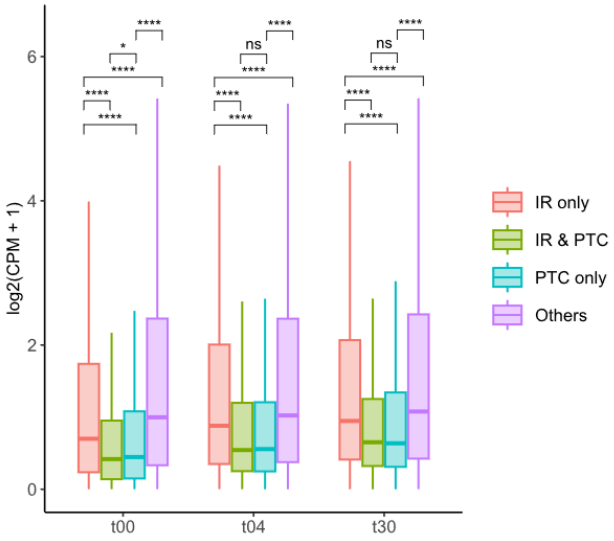
Supplementary Figure 10. Peptide confirmation of the translation of a novel exon in *EXOC1*. Related to Fig. 2. Mass spectrum of peptide VTEESVPSGENQSVTGGDEEVVDEYQELNAR, which confirms the translation of the detected *EXOC1* microexon. Matched b and y ions are highlighted. m/z, mass/charge ratio.



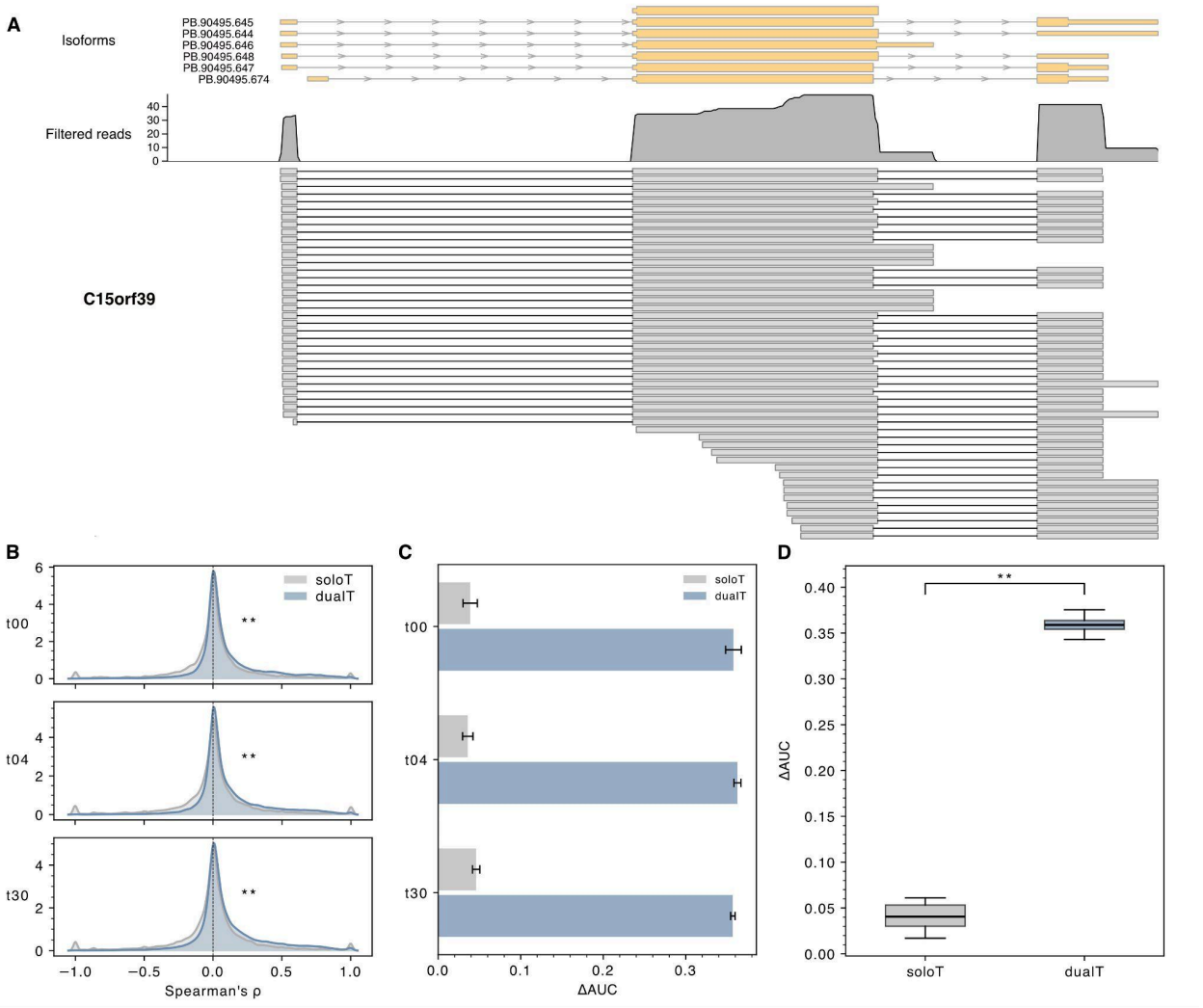
Supplementary Figure 11. River plot mapping mRNA trajectories (left) to protein trajectories (right) for the 8,493 genes with proteomic data. Source data are provided as a Source Data file.



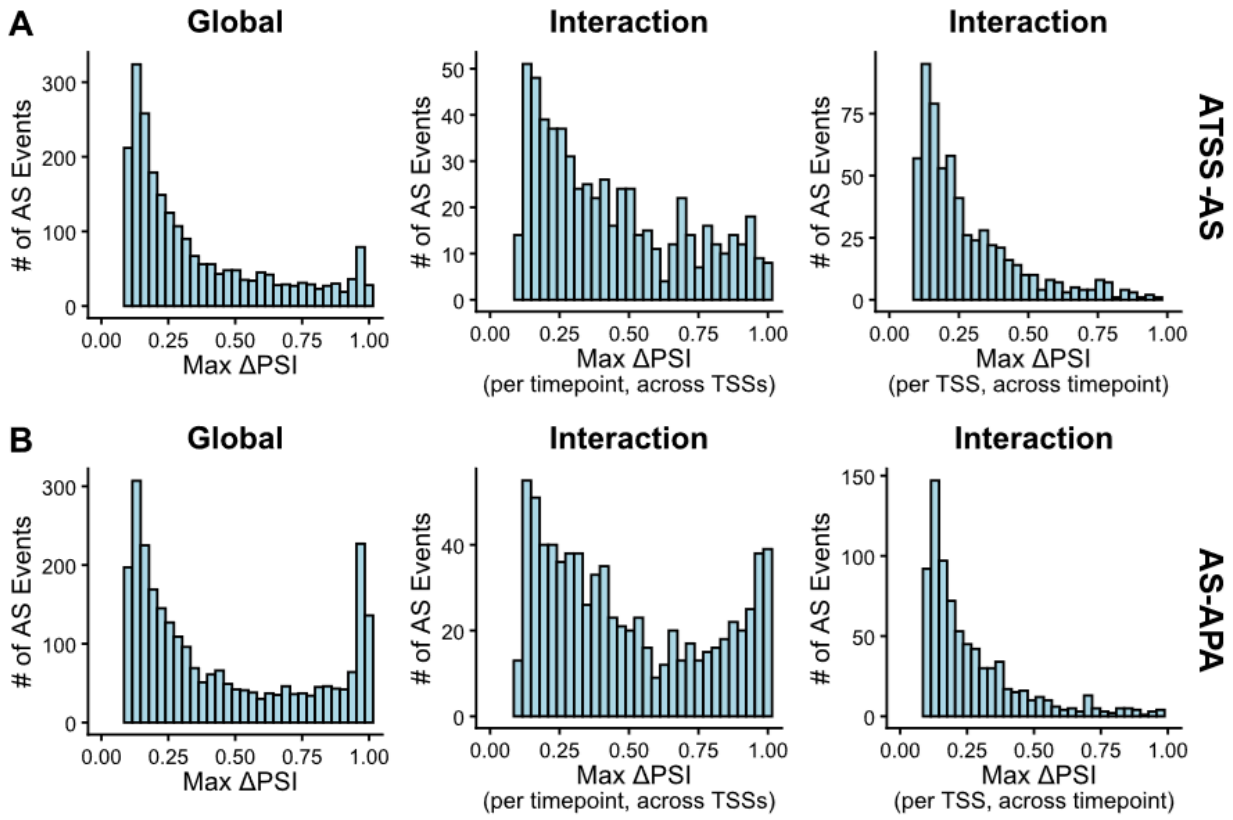
Supplementary Figure 12. Complete list of isoform switching events for alternative splicing and functional consequences analyzed across differentiation. Points represent the fraction of genes where switches primarily result in a given outcome (e.g., exon skipping). The size of the points represents the total number of genes with isoform switches resulting in either opposing consequence. Error bars represent 95% confidence intervals; significance determined by binomial test. Source data are provided as a Source Data file. ATTS = alternative transcription termination site; ATSS = alternative transcription start site; TSS = transcript start site; TTS = transcript termination site; ORF = open reading frame; NMD = nonsense-mediated decay; IDR = intrinsically disordered region.



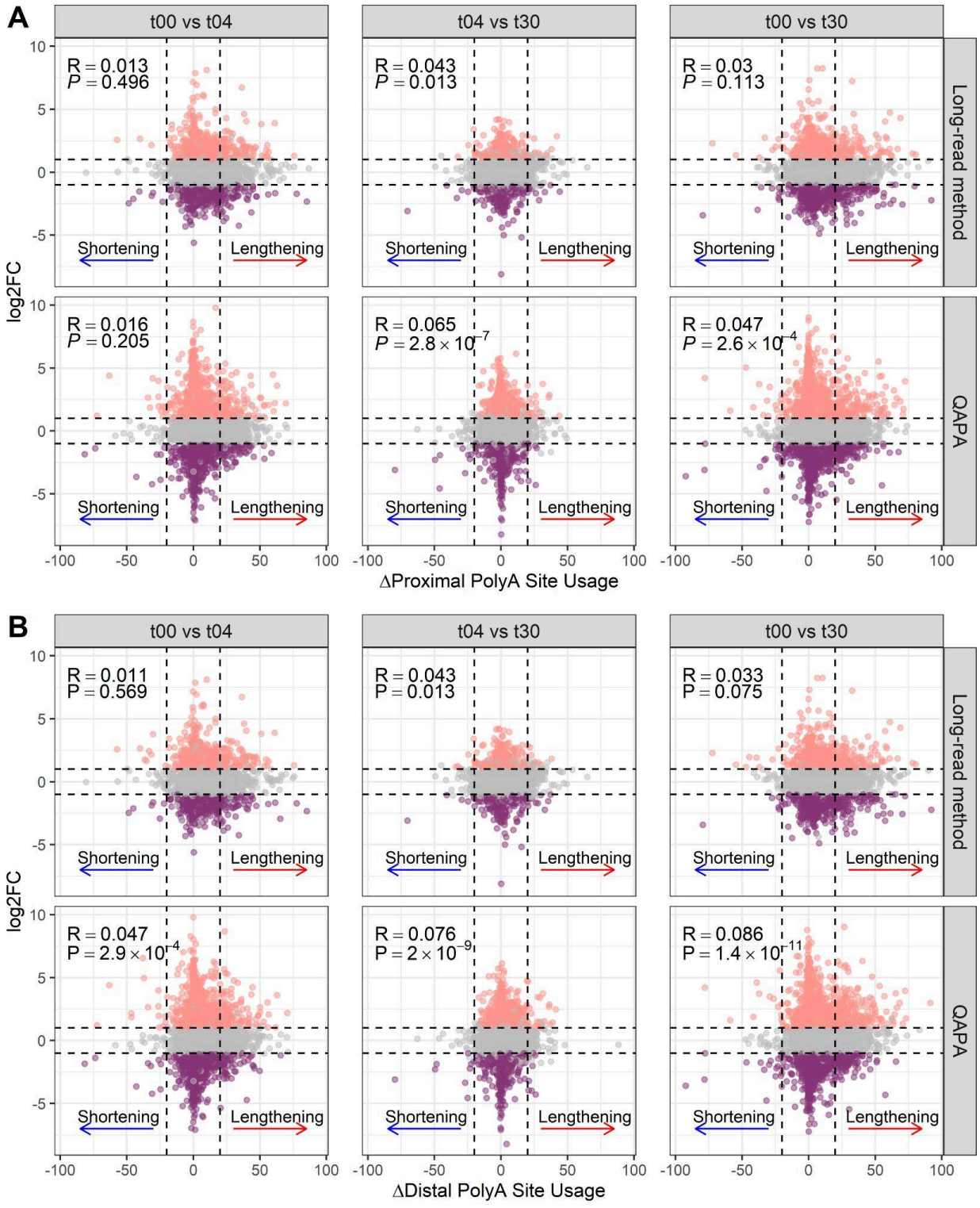
Supplementary Figure 13. Expression of transcripts grouped by intron retention and premature termination codons. Boxplots compare the expression of transcripts as having Intron Retention (IR) only, IR and a Premature Termination Codon (PTC), a PTC only, and all other transcripts. Statistical comparisons within each time point were performed using a Wilcoxon rank-sum test with Bonferroni correction (*P < 0.05; ****P < 0.0001; ns = not significant). Source data are provided as a Source Data file. IR = intron retention; PTC = premature termination codon.



Supplementary Figure 14. Positional coupling of TSS-PAS usage occurs within individual mRNA molecules. (A) Isoforms (top, orange), read coverage plot (middle, gray), a subset of long-read RNA-seq reads (bottom, gray with introns in thin black lines), from one single t30 sample. Distribution of Spearman's rho (B) and mean ΔAUC (C) for solo termini genes (soloT, exactly one unique CAGE peak or exactly one unique polyA site, gray) and dual alternative termini genes (dualT, more than one unique CAGE peak and more than one unique polyA site, blue) at t00, t04 and t30. K-S test, ** P-value < 10^{-16}. (D) Distribution of ΔAUC values across all 15 long-read RNA-seq samples for soloT (gray) and dualT genes (blue). t-test **P-value < 10^{-16}. Source data are provided as a Source Data file.

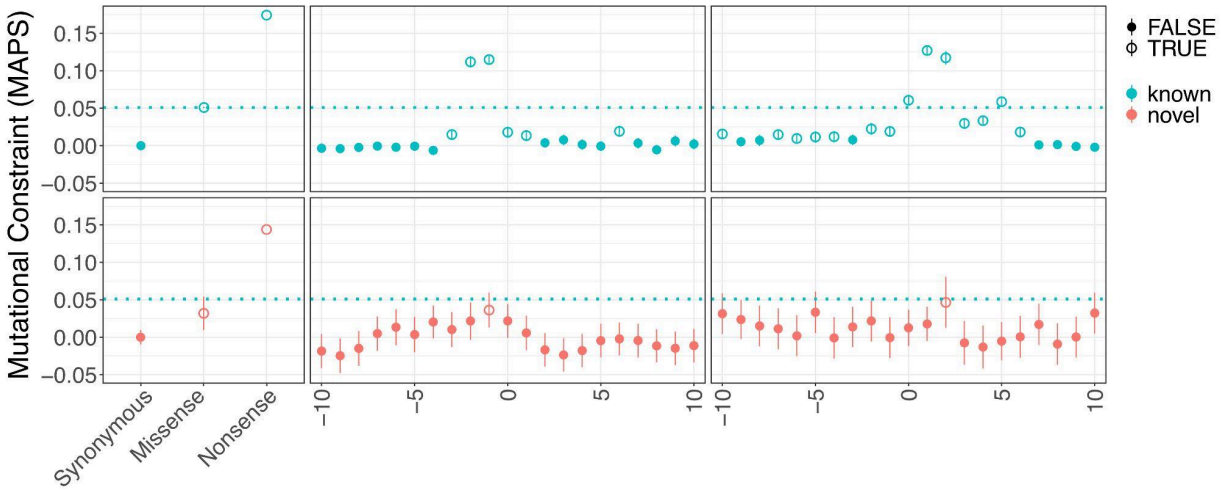


Supplementary Figure 15. (A) *Left*: The maximum ΔPSI (maximum PSI - minimum PSI) between TSSs, averaged across time, for all significant global ATSS-AS coordination events. *Middle*: The maximum change in PSI between TSSs at a single timepoint, for all significant interaction ATSS-AS coordination events. *Right*: The maximum change in PSI between timepoints, for a single TSS, for all significant interaction ATSS-AS coordination events. (**B)** *Left*: The maximum ΔPSI (maximum PSI - minimum PSI) between pA sites, averaged across time, for all significant global AS-APA coordination events. *Middle*: The maximum change in PSI between pA sites at a single timepoint, for all significant interaction AS-APA coordination events. *Right*: The maximum change in PSI between timepoints, for a single pA site, for all significant interaction AS-APA coordination events. Source data are provided as a Source Data file.

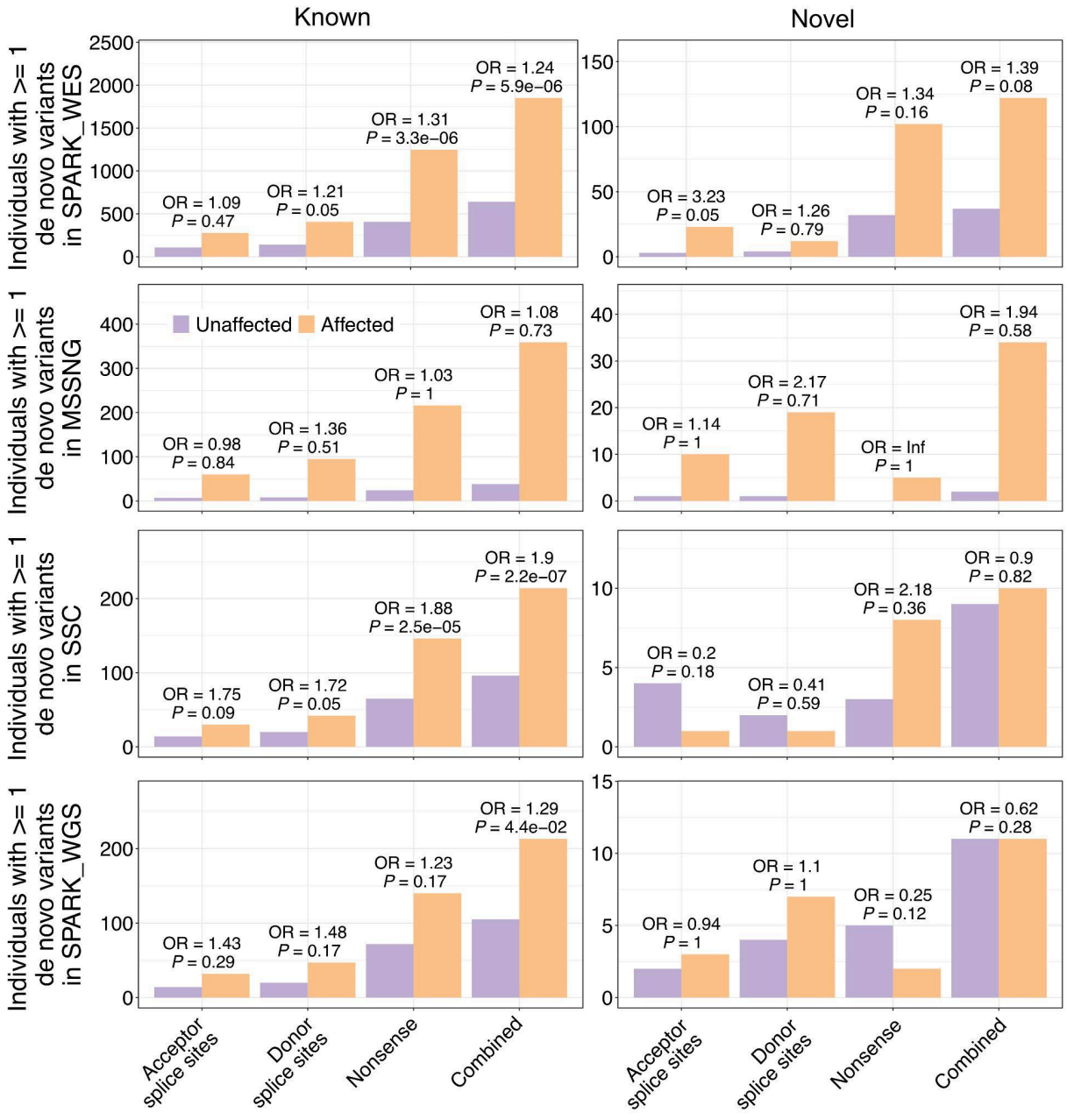


Supplementary Figure 16. Lack of correlation between 3′UTR length dynamics and mRNA expression changes. Scatter plots compare changes in gene expression (log_2FC, y-axis) against the change in (A) proximal polyA site usage (ΔPPAU, x-axis) and (B) distal

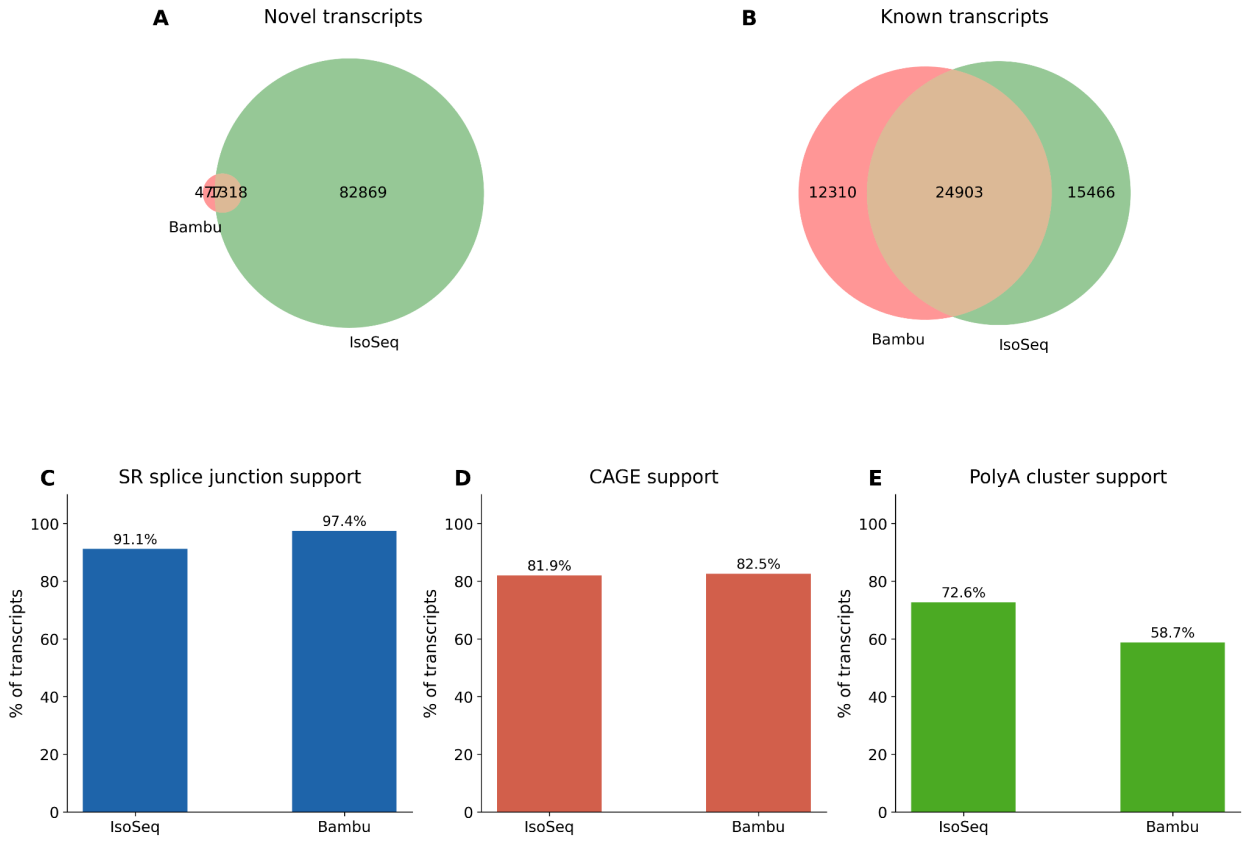
polyA site usage (ΔDPAU, x-axis) between pairs of differentiation timepoints. Top row uses our custom long-read analysis method and bottom row uses QAPA for validation. 3′UTR shortening (ΔPPAU/DPAU < − 20, blue arrows) and lengthening (ΔPPAU/DPAU > 20, red arrows) show no clear association with direction of gene expression changes ($|\log_2 FC| > 1$, FDR < 0.05), including gene upregulation (peach dots) or downregulation (purple dots). Pearson correlation coefficients (R) and *P*-values for all datapoints are shown in each panel. Vertical dotted lines indicate ΔPPAU/DPAU thresholds, and horizontal dotted lines indicate $\log_2 FC$ thresholds. Source data are provided as a Source Data file.



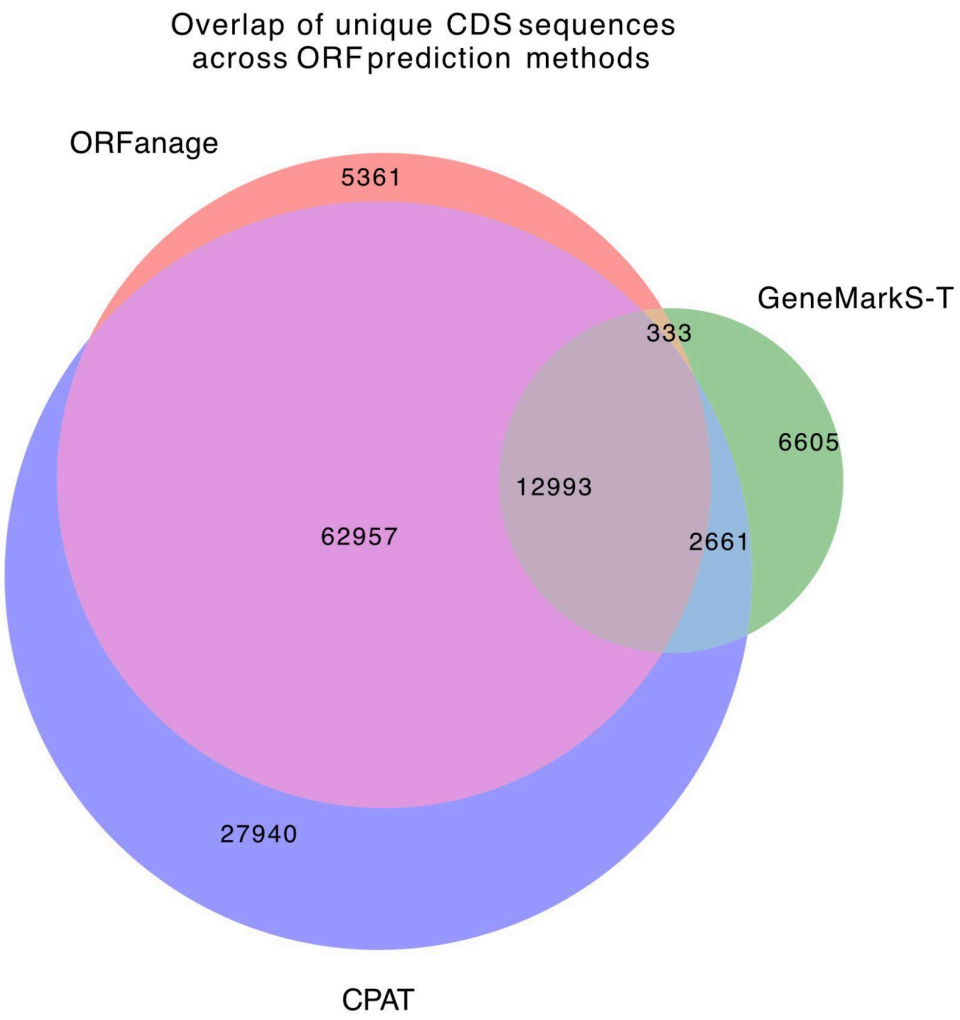
Supplementary Figure 17. Mutational constraint of coding and near-splice variants using whole-genome sequencing (WGS) data. Related to Figure 6. The plots show mutational constraint (MAPS scores) for single nucleotide variants from 76,215 individuals in the gnomAD v4.1 WGS cohort. The top panel shows constraint at Known (GENCODE V47) elements, and the bottom panel shows constraint at Novel elements identified in this study. For coding consequences (left), scores are shown for synonymous, missense, and nonsense variants. Positions surrounding splice acceptor and donor regions are also shown, with shaded areas highlighting the two essential splice sites. Open circles indicate variant classes or positions with mutational constraint significantly different from synonymous variants (FDR < 0.05, chi-squared test). Dashed horizontal lines indicate the MAPS scores for known missense and nonsense variants, provided for context. Error bars represent 95% CIs. Source data are provided as a Source Data file.



Supplementary Figure 18. Burden of disruptive *de novo* mutations in individual ASD cohorts. Related to Figure 6. These plots show the burden of *de novo* mutations in individuals affected by versus unaffected by ASD for known (left) and novel (right) genomic elements. Each row represents data from one ASD cohort. Odds ratios (OR) and *P*-values from a Fisher's exact test are shown. Source data are provided as a Source Data file.



Supplementary Figure 19. Comparison of Iso-seq and Bambu for transcript discovery. Venn diagrams showing the overlap in transcripts detected by Bambu and Iso-seq. Transcripts were compared by intron chain identity — two transcripts are considered equivalent if they share the same internal splice junctions, regardless of transcription start or end site variation. Only transcripts with more than 5 read counts in more than 2 samples were included. (A) Novel transcripts, defined as transcripts not matching any annotated isoform in the reference. (B) Known transcripts, defined as transcripts matching a reference isoform. Numbers indicate the count of transcripts unique to or shared between methods. (C) Percentage of transcripts from each long-read transcript discovery tool whose splice junctions are entirely supported by short-read RNA-seq (a transcript is considered supported only if every junction has ≥1 uniquely or multi-mapping read). (D) Percentage of transcripts whose 5′ end falls within 100 bp of a CAGE-seq peak (refTSS v3.3, hg38). (E) Percentage of transcripts whose 3′ end falls within a polyA cluster annotated in PolyASite v2.0 database. For all tools, only transcripts expressed at >5 reads in >2 samples are included. Numbers above bars indicate the percentage. Source data are provided as a Source Data file.



Supplementary Figure 20. Overlap of unique CDS sequences predicted by three ORF calling methods. Venn diagram showing the number of unique coding sequences shared between ORFanage (n = 81,644), GeneMarkS-T (n = 22,592), and CPAT (n = 106,551) across SFARI long-read transcripts. Sequences were deduplicated and compared by exact nucleotide identity. The majority of ORFanage predictions overlap with CPAT (62,957 shared exclusively between the two), while 12,993 sequences are recovered by all three methods. GeneMarkS-T shows the least overlap with the other tools, with 6,605 sequences unique to it.