A highly conserved program of neuronal microexons is misregulated in autistic brains

Manuel Irimia [1,2,9], Robert J. Weatheritt [1,3,10], Jonathan Ellis [1,10], Neelroop N. Parikshak [4,10], Thomas Gonatopoulos-Pournatzis [1], Mariana Babor [1], Mathieu Quesnel-Vallières [1], Javier Tapial [2], Bushra Raj [1], Dave O'Hanlon [1], Miriam Barrios-Rodiles [6], Michael J.E. Sternberg [5], Sabine P. Cordes [6,7], Frederick P. Roth[1,6,7,8], Jeffrey L. Wrana [6,7], Daniel H. Geschwind [4], and Benjamin J. Blencowe[1,7,9]

[1] Donnelly Centre, University of Toronto, Toronto, Canada
[2] EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), Barcelona, Spain
[3] MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, UK
[4] Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California
[5] Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London, UK
[6] Centre for Systems Biology, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada
[7] Department of Molecular Genetics, University of Toronto, Toronto, Canada
[8] Department of Computer Science, University of Toronto, Toronto, Canada
[9] Corresponding authors
[10] These authors contributed equally

Correspondence to:

Benjamin J. Blencowe, PhD and Manuel Irimia, PhD
Donnelly Centre,
University of Toronto
Toronto, ON, Canada
Office 416-978-3016; Fax 416-946-5545
Email: b.blencowe@utoronto.ca, mirimia@gmail.com

SUMMARY

Alternative splicing (AS) generates vast transcriptomic and proteomic complexity. However, which of the myriad of detected AS events provide important biological functions is not well understood. Here, we define the largest program of functionally coordinated, neural-regulated AS described to date in mammals. Relative to all other types of AS within this program, 3-15 nucleotide 'microexons' display the most striking evolutionary conservation and switch-like regulation. These microexons modulate the function of interaction domains of proteins involved in neurogenesis. Most neural microexons are regulated by the neuronal-specific splicing factor nSR100/SRRM4, through its binding to adjacent intronic enhancer motifs. Neural microexons are frequently misregulated in the brains of individuals with autism spectrum disorder, and this misregulation is associated with reduced levels of nSR100. The results thus reveal a highly conserved program of dynamic microexon regulation associated with the remodeling of protein interaction networks during neurogenesis, the misregulation of which is linked to autism.

INTRODUCTION

Alternative splicing (AS) – the process by which different pairs of splice sites are selected in precursor mRNA to generate multiple mRNA and protein products – is responsible for greatly expanding the functional and regulatory capacity of metazoan genomes (Braunschweig et al., 2013; Chen and Manley, 2009; Kalsotra and Cooper, 2011). For example, transcripts from over 95% of human multi-exon genes undergo AS, and most of the resulting mRNA splice variants are variably expressed between different cell and tissue types (Pan et al., 2008; Wang et al., 2008). However, the function of the vast majority of AS events detected to date are not known, and new landscapes of AS regulation remain to be discovered and characterized(Braunschweig et al., 2014; Eom et al., 2013). Moreover, since the misregulation of AS frequently causes or contributes to human disease, there is a pressing need to systematically define the functions of splice variants in disease contexts.

AS generates transcriptomic complexity through differential selection of cassette alternative exons, alternative 5´ and 3´ splice sites, mutually exclusive exons, and alternative intron retention. These events are regulated by the interplay of *cis*-acting motifs and *trans*-acting factors that control the assembly of spliceosomes (Chen and Manley, 2009; Wahl et al., 2009). The assembly of spliceosomes at 5´ and 3´ splice sites is typically regulated by RNA binding proteins (RBPs) that recognize proximal *cis*-elements, referred to as exonic/intronic splicing enhancers and silencers (Chen and Manley, 2009). An important advance that is facilitating a more general understanding of the role of individual AS events is the observation that many cell/tissue type- and developmentally-regulated AS events are coordinately controlled by individual RBPs, and that these events are significantly enriched in genes that operate in common biological processes and pathways (Calarco et al., 2011; Irimia and Blencowe, 2012; Licatalosi and Darnell, 2010).

AS can have dramatic consequences on protein function, and/or affect the expression, localization and stability of spliced mRNAs (Irimia and Blencowe, 2012). While cell and tissue differentially-regulated AS events are significantly under-represented in functionally defined, folded domains in proteins, they are enriched in regions of protein disorder that typically are surface accessible and embed short linear interaction motifs (Buljan et al., 2012; Ellis et al.,

2012; Romero et al., 2006). AS events located in these regions are predicted to participate in interactions with proteins and other ligands (Buljan et al., 2012; Weatheritt et al., 2012). Indeed, among a set of analyzed neural-specific exons enriched in disordered regions, approximately one third promoted or disrupted interactions with partner proteins (Ellis et al., 2012). These observations suggested that a widespread role for regulated exons is to specify cell and tissue type-specific protein interaction networks.

Human disease mutations often disrupt *cis*-elements that control splicing and result in aberrant AS patterns (Cartegni et al., 2002). Other disease changes affect the activity or expression of RBPs, causing entire programs of AS to be misregulated. For example, amyotrophic lateral sclerosis-causing mutations in the RBPs TLS/FUS and TDP43 affect AS and other aspects of post-transcriptional regulation (Polymenidou et al., 2012), and changes in the expression of the RBP RBFOX1 have been linked to misregulation of AS in the brains of individuals with autism spectrum disorder (ASD) (Voineagu et al., 2011). It is also widely established that misregulation of AS plays important roles in altering the growth and invasiveness of various cancers (David and Manley, 2010). As is the case with assessing the normal functions of AS, it is generally not known which disease-misregulated AS events cause or contribute to disease phenotypes.

Central to addressing the above questions is the importance of comprehensively defining AS programs associated with normal and disease biology. Gene prediction algorithms, high-throughput RNA sequencing (RNA-Seq) analysis methods, and RNA-Seq datasets generally lack the sensitivity and/or depth required to detect specific types of AS. In particular, microexons (Beachy et al., 1985; Coleman et al., 1987), defined here as 3-27 nucleotide (nt)-long exons, have been largely missed by genome annotations and transcriptome profiling studies (Volfovsky et al., 2003; Wu et al., 2013; Wu and Watanabe, 2005). This is especially true for microexons shorter than 15 nts. Furthermore, where alignment tools have been developed to capture microexons (Wu et al., 2013), they have not been applied to the analysis of different cell and tissue types, or disease states.

In this study, we developed a new RNA-Seq pipeline for the systematic discovery and analysis of all classes of AS, including microexons. By applying this pipeline to deep RNA-Seq datasets

from more than 50 diverse cell and tissue types, as well as developmental stages, from human and mouse, we define a large program of neural-regulated AS. Strikingly, neural-included microexons represent the most highly conserved and dynamically-regulated component of this program, and the corresponding genes are highly enriched in neuronal functions. These microexons are enriched on the surfaces of protein interaction domains and are under strong selection pressure to preserve reading frame. We also observe that microexons are frequently misregulated in the brains of autistic individuals, and that this misregulation is linked to the reduced expression of the neural-specific Ser/Arg-related splicing factor of 100kDa, nSR100/SRRM4. Collectively, our results reveal that alternative microexons represent the most highly conserved component of developmental AS regulation identified to date, and that they function in domain surface "microsurgery" to control interaction networks associated with neurogenesis. Microexons thus represent a new landscape for investigating the molecular consequences of AS (mis)regulation in nervous system development and ASD.

RESULTS

Global features of neural-regulated AS

An RNA-Seq analysis pipeline was developed to detect and quantify all AS event classes involving all hypothetically possible splice junctions formed by the usage of annotated and unannotated splice sites, including those that demarcate microexons. By applying this pipeline to more than 50 diverse cell and tissue types each from human and mouse (Table S1), we identified ~2,500 neural-regulated AS events in each species (Figure 1A and Table S2; Extended Experimental Procedures).

Nearly half of the neural-regulated AS events, including alternative retained introns, are predicted to generate protein isoforms both when the alternative sequence is included and skipped. In contrast, only ~20% of AS events not subject to neural regulation (hereafter 'non-neural' events) have the potential to generate alternative protein isoforms (Figure 1B; $p=2.7\times10^{-248}$, proportion test). Gene Ontology (GO) analysis shows that genes with neural-regulated AS events predicted to generate alternative protein isoforms form highly interconnected networks based on functions associated with neuronal biology, signaling pathways, structural components of the cytoskeleton

and the plasma membrane (Figure 1C). Consistent with previous results (Fagnani et al., 2007; Pan et al., 2004), there is little overlap (8.5%) between genes with neural-regulated AS and mRNA expression, although these subsets of genes are highly enriched in overlapping GO terms (40% in common; Figure S1). These data reveal the largest program of neural-regulated AS events defined to date, and that this program is associated with a broader range of functional processes and pathways linked to nervous system biology than previously detected (Boutz et al., 2007; Fagnani et al., 2007; Ule et al., 2005).

Highly conserved microexons are frequently neural specific

Further analysis of the neural-regulated AS program revealed a striking inverse relationship between the length of an alternative exon and its propensity to be specifically included in neural tissues. Increased neural-specific inclusion was detected for the majority of microexons (length ≤ 27 nt, Figure 2A); 60.7% of alternative microexons show increased neural 'percent spliced in' (PSI) (ΔPSI>15) versus 9.5% of longer (average ~135 nt) alternative exons ($p=1.9\times10^{-220}$, proportion test). This trend extends to microexons as short as 3 nt. RT-PCR validation experiments confirmed the RNA-Seq-detected regulatory profiles and inclusion levels of all (10/10) microexons analyzed across ten diverse tissues (R^2 = 0.92, n=107; Figure S2A). To further investigate the cell and tissue type specificity of microexon regulation, we used RNA-Seq data (Sofueva et al., 2013; Zhang et al., 2014; Zhang et al., 2013) to compare their inclusion levels in major glial cell types (astrocytes, microglia and oligodendrocytes), isolated neurons, and in muscle cells and tissues. While up to ~20% of the detected microexons showed increased PSIs in one or more glial cell types, and/or in muscle, compared to other non-neural tissues, the vast majority (>90%) of microexons display highest PSIs in neurons compared to all other cell and tissue types analyzed (Figure S2B-D, and Supplemental Information). These results indicate that tissue-regulated microexons are predominately neuronal-specific.

Relative to longer alternative exons, microexons, in particular those that are 3-15-nt long and neural-specifically included, are strongly enriched in multiple features indicative of functionally important AS. They are highly enriched for lengths that are multiples of three nts (Figure 2B), and a significantly larger fraction are predicted to generate alternative protein isoforms upon inclusion and exclusion, compared with longer exons (Figure 2C; $p<10^{-10}$, proportion test). They

are also significantly more often conserved at the levels of genomic sequence, detection in alternatively spliced transcripts, and neural-differential regulation (Figure 2D; Figure S2E, neural-regulated exons; $p<0.001$ for all pairwise comparisons, proportion tests. Similar results were obtained when comparing neural-regulated microexons and longer exons that have matching distributions of neural versus non-neural ΔPSI values (data not shown)). Of 308 neural-regulated microexons in human, 225 (73.5%) are neural-differentially spliced in mouse, compared to only 527 of 1390 (37.9%) longer neural-regulated exons. Remarkably, while microexons represent only ~1% of all AS events, they comprise approximately one third of all neural-regulated AS events conserved between human and mouse that are predicted to generate alternative protein isoforms (Figure S2F). Moreover, of ~150 analyzed mammalian, neural-regulated, 3-15-nt microexons, at least 55 are deeply conserved in vertebrate species spanning 400-450 million years of evolution, from zebrafish and/or shark to human (Table S3). This is in marked contrast to the generally low degree of evolutionary conservation of other types of AS across vertebrate species (Barbosa-Morais et al., 2012; Braunschweig et al., 2014; Merkin et al., 2012). Furthermore, comparable numbers of alternative microexons were detected in all analyzed vertebrate species, the majority of which are also strongly neural-specifically included (Figure 2E; Supplemental Information for details). Consistent with their striking regulatory conservation, sequences overlapping microexons, including both the upstream and downstream flanking intronic regions, are more highly conserved than sequences surrounding longer alternative exons (Figure 2F, S2G), including longer exons with a similar distribution of neural versus non-neural ΔPSI values (Figure S2H,I; data not shown).

Dynamic regulation of microexons during neuronal differentiation

To further investigate the functional significance of neural-regulated microexons, we used RNA-Seq data to analyze their regulation across six time points of differentiation of mouse embryonic stem (ES) cells into cortical glutamatergic neurons (Figure 3). Remarkably, of 219 neural-regulated microexons with sufficient read coverage across time points, 151 (69%) displayed a PSI switch ≥ 50 between ES cells and mature neurons, and 65 (30%) a switch of ≥ 90 (Figure 3). Unsupervised hierarchical clustering of PSI changes between consecutive time points (transitions T1 to T5) revealed several temporally-distinct regulatory patterns (Figure 3A). Most microexons show sharp PSI switches at late (T3 to T5) transitions during differentiation. These stages

correspond to maturing post-mitotic neurons when pan-neuronal markers are already expressed, and are subsequent to the expression of most neurogenic transcription factors (Figure S3A). This pattern of late activation (Figure S3B) suggests enrichment for important functions for microexons in terminal neurogenesis (Figure 1C). Despite the small number of genes representing clusters of kinetically-distinct sets of regulated microexons, each cluster revealed significant enrichment of specific GO terms including “regulation of GTPase activity” (Cluster I), “glutamate receptor binding” and “actin cytoskeleton organization” (Cluster V) (Table S4). These observations indicate that the dynamic switch-like regulation of microexons is intimately associated with the maturation of neurons.

The neural-specific splicing factor nSR100/SRRM4 regulates most neural microexons

Among several analyzed splicing regulators (Extended Experimental Procedures), knockdown and overexpression of nSR100 had the strongest effect on microexon regulation, with more than half of the profiled microexons displaying a pronounced change in inclusion level compared to controls (Figure 4A and S4A-S4H). Moreover, an analysis of RNA-Seq data from different neural cell types (Zhang et al., 2014) revealed that nSR100 has the strongest neuronal-specific expression relative to the other splicing regulators (Figure S4I, and data not shown), which is also consistent with its immunohistochemical detection in neurons but not glia (Calarco et al., 2009). Recently, we have shown that nSR100 promotes the inclusion of a subset of (longer) neural exons via binding to intronic UGC motifs proximal to suboptimal 3´ splice sites (Raj et al., 2014). Consistent with these results, and supporting a direct role for nSR100 in microexon regulation, RNA sequence tags cross-linked to nSR100 in vivo are also highly enriched in intronic sequences containing UGC motifs, located adjacent to the 3´ splice sites of nSR100-regulated microexons (Figure 4B, C; $p<0.0001$ for all comparisons, Wilcoxon Rank Sum test). Relative to longer exons, we additionally observe that neural-regulated microexons are associated with weak 3´ splice sites and strong 5´ splice sites (Figure S4J). nSR100 thus has a direct and extensive role in the regulation of the neural microexon program.

Distinct protein regulatory properties of microexons

Neural-regulated microexons, in particular those that are 3-15-nt long, possess multiple properties that distinguish them from longer neural-regulated exons (Figures 5 and S5). A significantly

smaller fraction overlap predicted disordered amino acid residues (Figures 5A and S5A-D; $p < 1.3\times10^{-4}$; 3-way Fisher Exact tests), whereas a significantly higher fraction overlap modular protein domains (Figures 5B and S5E; ~2-fold increase, $p=1.0\times10^{-54}$, proportion test). In contrast, microexon residues overlapping protein domains are significantly more often surface-accessible and enriched in charged residues (Figures 5C,D, and S5F-I; $p<10^{-7}$ for all comparisons, proportion test) than are residues overlapping longer neural or non-neural exons. Moreover, when not overlapping protein domains, microexons are significantly more often located immediately adjacent (i.e. within 5 amino acids) to folded protein domains (Figures 5E and S5J,K). These results suggest that a common function of microexons may be to modulate the activity of overlapping or adjacent protein domains. Supporting this view, among 49 available and *de novo*-modeled tertiary protein structures containing microexons, the corresponding residues are largely surface accessible amd unlikely to significantly affect the folding of the overlapping or adjacent protein domains (Figure S6A; Extended Experimental Procedures).

Microexons modulate the function of interaction domains

Neural-regulated microexons are significantly enriched in domains that function in peptide and lipid-binding interactions (Figures 5F and S5L; $p = 1.7\times10^{-6}$, proportion test). Overall, genes with microexons are highly enriched in modular domains involved in cellular signaling, such as SH3 and PH domains (Figure S5M). Conversely, unlike longer neural exons (Buljan et al., 2012; Ellis et al., 2012), they are depleted of linear binding motifs (Figure 5G and S5N, $p<0.005$, proportion tests for all comparisons). Moreover, proteins containing microexons are significantly more often central in protein-protein interaction networks and detected in protein complexes compared to proteins with other types of alternative exons (Figures 5H and S5O,P, $p\leq0.004$ for all comparisons, Wilcoxon Rank Sum test). Taken together with the data in Figure 1, these results suggest that microexons may often regulate interaction domains to facilitate the remodeling of protein interaction networks associated with signaling and other aspects of neuronal maturation and function.

To test this hypothesis, we employed luminescence-based mammalian interactome mapping (LUMIER; (Barrios-Rodiles et al., 2005; Ellis et al., 2012)) and co-immunoprecipitation-western blot assays to investigate whether the insertion of a highly conserved, neural-regulated 6-nt

microexon in the nuclear adaptor Apbb1 affects its known interactions with the histone acetyltransferase Kat5/Tip60, and amyloid precursor protein App (Figure 6A-D). Previous genetic and functional studies have revealed multiple functions for the Apbb1-Kat5 complex (Cao and Sudhoff, 2001; Stante et al., 2009), and that the loss of Kat5 activity is associated with developmental defects that impact learning and memory (Pirooznia et al., 2012; Wang et al., 2004; Wang et al., 2009) (see Discussion). Apbb1 contains two phosphotyrosine binding domains, PTB1 and PTB2, which bind Kat5 and App, respectively (Cao and Sudhoff, 2001). Exemplifying the distinct protein features of neural microexons described above (Figure 5), the Apbb1 microexon adds two charged residues (Arg and Glu) to the PTB1 domain near its predicted interaction surface (Figures 6A,B; Extended Experimental Procedures). LUMIER and co-immunoprecipitation-western analysis reveals that inclusion of the microexon significantly enhances the interaction with Kat5, whereas there is little to no effect on the interaction with App (Figures 6C,D and Figure S6B,C). Substitution of both microexon residues with alanine also enhanced the Kat5 interaction, although to a lesser extent than the presence of Arg and Glu (Figure 6C). This suggests that the primary function of this microexon is to extend the interface with which Apbb1 binds its partner proteins.

We also examined the function of a 9-nt microexon in the AP1S2 subunit of the adaptor-related protein complex 1 (AP1). The AP1 complex functions in the intracellular transport of cargo proteins between the *trans*-Golgi apparatus and endosomes by linking clathrin to the cargo proteins during vesicle membrane formation (Kirchhausen, 2000), and is important for the somatodendritic transport of proteins required for neuronal polarity (Farias et al., 2012). Interestingly, mutations in AP1S2 have been previously implicated in phenotypic features associated with ASD and X-linked mental retardation (Borck et al., 2008; Tarpey et al., 2006). Co-immunoprecipitation-western analyses reveal that the microexon in AP1S2 strongly promotes its interaction with another AP1 subunit, AP1B1 (Figure 6E and S6D). This observation thus provides additional evidence supporting an important role for microexons in the control of protein interactions that function in neurons.

Microexons are misregulated in individuals with Autism Spectrum Disorder

The properties of microexons described above suggest that their misregulation could be associated with neurological disorders. To investigate this possibility, we analyzed RNA-Seq data from the superior temporal gyrus (Brodmann areas ba41/42/22) from post-mortem samples from individuals with ASD and control subjects, matched for age, gender and other variables (Experimental Procedures). These samples were stratified based on the strength of an ASD-associated gene expression signature (Voineagu et al., 2011), and subsets of 12 ASD samples with the strongest ASD-associated differential gene expression signatures and 12 controls were selected for further analysis. Remarkably, within these samples, 126 of 504 (30%) detected alternative microexons display a mean ΔPSI > 10 between ASD and control subjects (Figure 7A), of which 113 (90%) also display neural-differential regulation. By contrast, only 825 of 15,405 (5.4%) longer (i.e. >27 nt) exons show such misregulation (Figure 7A), of which 285 (35%) correspond to neural-regulated exons. Significant enrichment for misregulation among microexons compared to longer exons was also observed when restricting the analysis to neural-regulated exons, including subsets of neural-regulated microexons and longer exons with similar distributions of neural versus non-neural ΔPSI values (Figure S7A; $p<2\times10^{-4}$, proportion test; data not shown). Similar results were observed when analyzing data from a different brain region (Brodmann area ba9) from the same individuals (data not shown). RT-PCR experiments on a representative subset of profiled tissues confirmed increased misregulation of microexons in autistic versus control brain samples (Figure S7B). Analysis of the proportions of microexons displaying coincident misregulation revealed that the vast majority (81.3%) have a ΔPSI>10 in at least half of the ASD stratified brain samples (Figure S7C). However, only 26.9% (32/119) of the genes containing misregulated microexons overlapped with the 2,519 genes with significant ASD-associated misregulation at the level of gene expression. This reveals that largely distinct subsets of genes are misregulated at the levels of expression and microexon splicing in the analyzed ASD subjects. In contrast, a comparison of autistic subjects that possessed a weaker ASD-related differential gene expression signature did not reveal significant misregulation of microexons, or of longer exons (data not shown). These data reveal frequent misregulation of microexon splicing in the brain cortices of some individuals with ASD.

Consistent with a widespread and important role for nSR100 in the regulation of microexons (Figure 4), nSR100 mRNA expression is, on average, significantly downregulated in the brains of

the analyzed ASD versus control subjects, and to an even greater extent in brain samples with the strongest ASD-associated signature compared to the controls (~10%, p=0.014, FDR<0.1, Figure 7B and data not shown). These differences were confirmed by qRT-PCR assays for a representative subset of individuals ($p<2.8\times10^{-4}$ for all normalizations; two-sided T-test; Figure S7D). Moreover, relative to other exons, nSR100-dependent microexons are significantly more often misregulated in brain tissues from ASD compared to control subjects (Figure 7C; p<0.01 for all comparisons, proportion test). Notably, we also observe significantly higher correlations between microexon inclusion and nSR100 mRNA expression levels across the stratified ASD samples and controls, for those microexons regulated by nSR100 relative to those microexons that are not regulated by this factor (Figure 7D; $p=1.4\times10^{-7}$, Wilcoxon Sum Rank test).

A GO analysis of genes with ASD-associated misregulation of microexons reveals significant enrichment of terms related to axonogenesis and synapse biology (Figure 7E), processes that have been previously implicated in autism (Gilman et al., 2011; Parikshak et al., 2013; Voineagu et al., 2011). Many of the corresponding genes act in common pathways and/or physically interact through protein-protein interactions (Figure 7F). Moreover, misregulated microexons are also significantly enriched in genes that have been genetically linked to ASD (p < 0.0005, Fisher exact test), including many relatively well-established examples such as *DNTA*, *ANK2*, *ROBO1*, *SHANK2*, *AP1S2*. Other genes with misregulated microexons have been linked to learning or intellectual disability (e.g. *APBB1*, *TRAPPC9*, *RAB3GAP1*). In this regard, it is interesting to note that the microexons we have analyzed in APBB1 and AP1S2 are significantly misregulated in the brain samples from ASD subjects (p<0.05 Wilcoxon Sum Rank test; Figure S7E). Taken together with data in Figures 5 and 6, the results suggest that the misregulation of microexons, as well as of longer alternative exons (Corominas et al., 2014; Voineagu et al., 2011), may impact protein interaction networks that are required for normal neuronal development and synaptic function. Disruption of microexon-regulated protein interaction networks is therefore a potentially important mechanism underlying ASD and likely other neurodevelopmental disorders.

DISCUSSION

In this study, we show that alternative microexons display the highest degrees of genomic sequence conservation, tissue-specific regulatory conservation, and frame-preservation potential, relative to all other classes of AS detected to date in vertebrate species. Unlike longer neural-regulated exons, neural microexons are significantly enriched in surface-accessible, charged amino acids that overlap or lie in close proximity to protein domains, including those that bind linear motifs. Together with their remarkably dynamic regulation, these observations suggest that microexons contribute important and complementary roles to longer neural exons in the remodeling of protein interaction networks that operate during neuronal maturation.

Most microexons display high inclusion at late stages of neuronal differentiation in genes (e.g. *Src* (Black, 1991), *Bin1*, *Agrn*, *Dock9*, *Shank2*, *Robo1*) associated with axonogenesis and the formation and function of synapses. Supporting such functions, an alternative microexon overlapping the SH3A domain of Intersectin 1 (Itsn1) has been reported to promote an interaction with Dynamin 1, and was proposed to modulate roles of Itsn1 in endocytosis, cell signaling and/or actin-cytoskeleton dynamics (Dergai et al., 2010). A neural-specific microexon in Protrudin/Zfyve27 was recently shown to increase its interaction with the vesicle-associated membrane protein-associated protein (VAP), and to promote neurite outgrowth (Ohnishi et al., 2014). Similarly, in the present study, we show that a 6 nt neural microexon in Apbb1/Fe65 promotes an interaction with Kat5/Tip60. Apbb1 is an adapter protein that functions in neurite outgrowth (Cheung et al., 2014; Ikin et al., 2007) and synaptic plasticity (Sabo et al., 2003), processes that have been linked to neurological disorders including ASD (Hussman et al., 2011). Consistent with these findings, we have previously shown that nSR100 promotes neurite outgrowth (Calarco et al., 2009). In the present study we further demonstrate that it controls the switch-like regulation of most neural microexons, and that its reduced expression is linked to the altered splicing of microexons in the brains of subjects with ASD.

Many of the conserved, neural-regulated microexons identified in this study are misregulated in ASD individuals, including the microexon in AP1S2 that strongly promotes an interaction with the AP1B1 subunit of the AP1 intracellular transport complex. Intriguingly, several other genes containing microexons are genetically linked to ASD, intellectual disability and/or functions in memory and learning (see Results). Another link to ASD is the observation that nSR100 is

strongly co-expressed in the developing human brain in a gene network module, M2, which is enriched for rare *de novo* ASD-associated mutations (Parikshak et al., 2013). Furthermore, additional genes containing microexons may have as yet undiscovered roles in ASD and or other neuropsychiatric disorders. For example, the microexon in APBB1 is also significantly misregulated in brain tissues from ASD subjects (Figure S7B,E). It is possible that the misregulation of microexons, at least in part through altered expression of nSR100, perturbs protein interaction networks required for proper neuronal maturation and function, thus contributing to ASD as well as other neurodevelopmental disorders. Consistent with this view, recent reports have begun to link individual microexons with neurodevelopmental disorders, including ASD (Zhu et al., 2014), schizophrenia (Ovadia and Shifman, 2011) and epilepsy (Rusconi et al., 2014). The discovery and characterization of widespread, neural-regulated microexons in the present study thus enables a systematic investigation of new and highly conserved mechanisms controlling protein interaction networks associated with vertebrate nervous system development and neurological disorders.

EXPERIMENTAL PROCEDURES

RNA-Seq data and genomes

Unless stated otherwise, RNA-Seq data was generated from Poly(A)$^+$ RNA (Table S1). Analyses used the following genome releases: *Homo sapiens*, hg19, *Mus musculus*, mm9; *Gallus gallus*, galGal3; *Xenopus tropicalis*, xenTro3; *Danio rerio*, danRer7; *Callorhinchus milii*, v1.0).

Alternative splicing analysis pipeline

A multi-module analysis pipeline was developed that uses RNA-Seq, EST and cDNA data, as well as gene annotations and evolutionary conservation, to assemble libraries of exon-exon-junctions (EEJs) for subsequent read alignment to detect and quantify AS events in RNA-Seq data. For cassette exons, three complementary modules were developed for assembling EEJs: (i) A “transcript-based module”, employing cufflinks (Trapnell et al., 2010) and alignments of ESTs and cDNAs with genomic sequence (Khare et al., 2012)); (ii) A “splice site-based module”, utilizing joining of all hypothetically-possible EEJ combinations from annotated and *de novo* splice sites; and (3) A “microexon module”, including *de novo* searching of pairs of donor and

acceptor splice sites in intronic sequence. Alt3 or Alt5 events were quantified based on the fraction of reads supporting the usage of each alternative splice site. Intron retention was analyzed as recently described (Braunschweig et al., 2014). See Extended Experimental Procedures for additional details.

LUMIER assay

HEK-293T cells were transiently transfected using Polyfect (Qiagen) with *Renilla* Luciferase (RL)-tagged Apbb1, with or without inclusion of the microexon, or with a version consisting of two alanine substitutions, together with 3Flag-tagged Kat5. Subsequent steps were performed essentially as described previously (Ellis et al., 2012).

Immunoprecipitation and immunoblotting

HEK-293T cells were transiently transfected using Lipofectamine 2000 (Life Technologies). Cells were lysed in 0.5% TNTE. After pre-clearing with protein G-Sepharose, lysates were incubated with anti-Flag M2 antibody (Sigma) or anti-Hemagglutinin (HA)-antibody (Roche) bound to Protein-G Dynabeads (Life Technologies^{TM}) for 2 hours at 4℃. Immunoprecipitates were washed 5 times with 0.1% TNTE, subjected to SDS-PAGE, transferred onto nitrocellulose and immunoblotted with the anti-Hemagglutinin (HA)-antibody (Roche) or anti-Flag M2 antibody (Sigma). Detection was achieved using horseradish peroxidase-conjugated rabbit anti-rat (Sigma) or sheep anti-mouse secondary antibodies (GE Healthcare) and chemiluminescence. ImageJ was used for quantification of band intensities.

Analysis of microexon regulation

Available RNA-Seq data from splicing factor-deficient or -overexpressing systems were used to identify misregulated exons and microexons (see Extended Experimental Procedures). To investigate regulation by nSR100, we used PAR-iCLIP data and motif enrichments analyses, as recently described (Raj et al., 2014).

Comparison of ASD and control brain samples

We analyzed 24 autistic individuals and 24 controls matched by age and gender. Samples from superior temporal gyrus (Brodmann areas ba41/42/22) were dissected retaining grey matter from

all cortical layers, and RNA was isolated using the miRNeasy kit (Qiagen). Ribosomal RNA was depleted from 2ug total RNA with the Ribo-Zero Gold kit (Epicentre), and then size-selected with AMPure XP beads (Beckman Coulter). An average of 64 million, 50bp paired-end reads were generated for each sample (Table S1). The 12 samples with the strongest ASD-associated differential gene expression signature, and 12 control samples with a signal that is closest to the median of all controls, were selected for downstream analyses (Extended Experimental Procedures for details). Sample selection was independent of any information on splicing changes.

ACKNOWLEDGEMENTS

The authors thank the Eunice Kennedy Shriver NICHD Brain and Tissue Bank for Developmental Disorders, the Autism Tissue Program, and the Harvard Brain Tissue Resource Center for providing brain samples. We also thank Ulrich Braunschweig for assistance with CLIP-Seq analyses, Benjamin Lang for advice on surface accessibility measurements, Nuno Barbosa-Morais for guidance on statistical testing, and Serge Gueroussov and Jonathan Roth for helpful discussions and comments on the manuscript. MI holds a LTF from the Human Frontiers Science Program Organization. RJW holds a Canadian Institute of Health Research (CIHR) Postdoctoral Fellowship. NNP holds an NIMH NRSA fellowship. MB is supported by a fellowship from the Department of Cell and Systems Biology, University of Toronto. MQV holds a Banting and Best CIHR Scholarship. TGP is supported by fellowships from EMBO and OSCI. This research was supported by grants from the CIHR (BJB, JLW, FPR), Ontario Research Fund (JLW, BJB and others), Alzheimer's Research Foundation (BJB), University of Toronto McLaughlin Centre (BJB), NIH/NHGRI (FPR, P50 HG004233), NIMH (DHG, 5R37MH060233 and 5R01MH094714), and the Simons Foundation (DHG, SFARI 206744). BJB holds the Banbury Chair of Medical Research at the University of Toronto.

REFERENCES

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.* (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587-1593.

Barrios-Rodiles, M., Brown, K.R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R.S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I.W., *et al.* (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. Science *307*, 1621-1625.

Beachy, P.A., Helfand, S.L., and Hogness, D.S. (1985). Segmental distribution of bithorax complex proteins during Drosophila development. Nature *313*, 545-551.

Black, D.L. (1991). Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? Genes Dev *5*, 389-402.

Borck, G., Molla-Herman, A., Boddaert, N., Encha-Razavi, F., Philippe, A., Robel, L., Desguerre, I., Brunelle, F., Benmerah, A., Munnich, A., *et al.* (2008). Clinical, cellular, and neuropathological consequences of AP1S2 mutations: further delineation of a recognizable X-linked mental retardation syndrome. Hum Mutat *29*, 966-974.

Boutz, P.L., Stoilov, P., Li, Q., Lin, C.H., Chawla, G., Ostrow, K., Shiue, L., Ares, M.J., and Black, D.L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev *21*, 1636-1652.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Frey, B.J., Irimia, M., and Blencowe, B.J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res, pii: gr.177790.177114.

Braunschweig, U., Gueroussov, S., Plocik, A.M., Graveley, B.R., and Blencowe, B.J. (2013). Dynamic integration of splicing within gene regulatory pathways. Cell *152*, 1252-1269.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell *46*, 871-883.

Calarco, J.A., Superina, S., O'Hanlon, D., Gabut, M., Raj, B., Pan, Q., Skalska, U., Clarke, L., Gelinas, D., van der Kooy, D., *et al.* (2009). Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. Cell *138*, 898-910.

Calarco, J.A., Zhen, M., and Blencowe, B.J. (2011). Networking in a global world: establishing functional connections between neural splicing regulators and their target transcripts. RNA *17*, 775-791.

Cao, X., and Sudhoff, T.C. (2001). A transcriptionally active complex of APP with Fe65 and histone acetyltransferase Tip60. Science *293*, 115-120.

Cartegni, L., Chew, S.L., and Krainer, A.R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet *3*, 285-298.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol *10*, 741-754.

Cheung, H.N., Dunbar, C., Morotz, G.M., Cheng, W.H., Chan, H.Y., Miller, C.C., and Lau, K.F. (2014). FE65 interacts with ADP-ribosylation factor 6 to promote neurite outgrowth. FASEB J *28*, 337-349.

Coleman, K.G., Poole, S.J., Weir, M.P., Soeller, W.C., and Kornberg, T. (1987). The invected gene of Drosophila: sequence analysis and expression studies reveal a close kinship to the engrailed gene. Genes Dev *1*, 19-28.

Corominas, R., Yang, X., Lin, G.N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S.A., *et al.* (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat Commun *5*, 3650.

David, C.J., and Manley, J.L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev *24*, 2343-2364.

Dergai, M., Tsyba, L., Dergai, O., Zlatskii, I., Skrypkina, I., Kovalenko, V., and Rynditch, A. (2010). Microexon-based regulation of ITSN1 and Src SH3 domains specificity relies on introduction of charged amino acids into the interaction interface. Biochem Biophys Res Commun *399*, 307-312.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., *et al.* (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell *46*, 884-892.

Eom, T., Zhang, C., Wang, H., Lay, K., Fak, J., Noebels, J.L., and Darnell, R.B. (2013). NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. Elife *2*, e00178.

Fagnani, M., Barash, Y., Ip, J., Misquitta, C., Pan, Q., Saltzman, A., Shai, O., Lee, L., Rozenhek, A., Mohammad, N., *et al.* (2007). Functional coordination of alternative splicing in the mammalian central nervous system. Genome Biology *8*, R108.

Farias, G.G., Cuitino, L., Guo, X., Ren, X., Jarnik, M., Mattera, R., and Bonifacino, J.S. (2012). Signal-mediated, AP-1/clathrin-dependent sorting of transmembrane receptors to the somatodendritic domain of hippocampal neurons. Neuron *75*, 810-823.

Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron *70*, 898-907.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., *et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068-1081.

Hubbard, K.S., Gut, I.M., Lyman, M.E., and McNutt, P.M. (2013). Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. F1000Res *2*, 35.

Hussman, J.P., Chung, R.H., Griswold, A.J., Jaworski, J.M., Salyakina, D., Ma, D., Konidari, I., Whitehead, P.L., Vance, J.M., Martin, E.R., *et al.* (2011). A noise-reduction GWAS analysis implicates altered regulation of neurite outgrowth and guidance in autism. Mol Autism *2*, 1.

Ikin, A.F., Sabo, S.L., Lanier, L.M., and Buxbaum, J.D. (2007). A macromolecular complex involving the amyloid precursor protein (APP) and the cytosolic adapter FE65 is a negative regulator of axon branching. Mol Cell Neurosci *35*, 57-63.

Irimia, M., and Blencowe, B.J. (2012). Alternative splicing: decoding an expansive regulatory layer. Curr Opin Cell Biol *24*, 323-332.

Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet *12*, 715-729.

Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M., *et al.* (2012). 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. Nat Struct Mol Biol *19*, 1037-1043.

Kirchhausen, T. (2000). Clathrin. Annu Rev Biochem *69*, 699-727.

Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat Rev Genet *11*, 75-87.

Merkin, J., Russell, C.B., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593-1599.

Ohnishi, T., Shirane, M., Hashimoto, Y., Saita, S., and Nakayama, K.I. (2014). Identification and characterization of a neuron-specific isoform of protrudin. Genes Cells *19*, 97-111.

Ovadia, G., and Shifman, S. (2011). The genetic variation of RELN expression in schizophrenia and bipolar disorder. PLoS One *6*, e19955.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40*, 1413-1415.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., *et al.* (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell *16*, 929-941.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell *155*, 1008-1021.

Pirooznia, S.K., Sarthi, J., Johnson, A.A., Toth, M.S., Chiu, K., Koduri, S., and Elefant, F. (2012). Tip60 HAT activity mediates APP induced lethality and apoptotic cell death in the CNS of a Drosophila Alzheimer's disease model. PLoS One *7*, e41776.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Bennett, C.F., Cleveland, D.W., and Yeo, G.W. (2012). Misregulated RNA processing in amyotrophic lateral sclerosis. Brain Res *1462*, 3-15.

Raj, B., Irimia, M., Braunschweig, U., Sterne-Weiler, T., O'Hanlon, D., Yuan-Lin, Z., Chen, I.G., Easton, L., Ule, J., Gingras, A.C., *et al.* (2014). Global regulatory mechanism underlying the activation of an exon network required for neurogenesis. Mol Cell *56*, 90-103.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., *et al.* (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci USA *103*, 8390-8395.

Rusconi, F., Paganini, L., Braida, D., Ponzoni, L., Toffolo, E., Maroli, A., Landsberger, N., Bedogni, F., Turco, E., Pattini, L., *et al.* (2014). LSD1 Neurospecific Alternative Splicing Controls Neuronal Excitability in Mouse Models of Epilepsy. Cereb Cortex *In press*.

Sabo, S.L., Ikin, A.F., Buxbaum, J.D., and Greengard, P. (2003). The amyloid precursor protein and its regulatory protein, FE65, in growth cones and synapses in vitro and in vivo. J Neurosci *23*, 5407-5415.

Sofueva, S., Yaffe, E., Chan, W.C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J *32*, 3119-3129.

Stante, M., Minopoli, G., Passaro, F., Raia, M., Vecchio, L.D., and Russo, T. (2009). Fe65 is required for Tip60-directed histone H4 acetylation at DNA strand breaks. Proc Natl Acad Sci USA *106*, 5093-5098.

Tarpey, P.S., Stevens, C., Teague, J., Edkins, S., O'Meara, S., Avis, T., Barthorpe, S., Buck, G., Butler, A., Cole, J., *et al.* (2006). Mutations in the gene encoding the Sigma 2 subunit of the adaptor protein 1 complex, AP1S2, cause X-linked mental retardation. Am J Hum Genet *79*, 1119-1124.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van, B.a., M.J, Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.* (2005). Nova regulates brain-specific splicing to shape the synapse. Nat Genet *37*, 844-852.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380-384.

Volfovsky, N., Haas, B.J., and Salzberg, S.L. (2003). Computational discovery of internal micro-exons. Genome Res *13*, 1216-1221.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

Wang, B., Hu, Q., Hearn, M.G., Shimizu, K., Ware, C.B., Liggitt, D.H., Jin, L.W., Cool, B.H., Storm, D.R., and Martin, G.M. (2004). Isoform-specific knockout of FE65 leads to impaired learning and memory. J Neurosci Res *75*, 12-24.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, Y., Zhang, M., Moon, C., Hu, Q., Wang, B., Martin, G., Sun, Z., and Wang, H. (2009). The APP-interacting protein FE65 is required for hippocampus-dependent learning and long-term potentiation. Learn Mem *16*, 537-544.

Weatheritt, R.J., Davey, N.E., and Gibson, T.J. (2012). Linear motifs confer functional diversity onto splice variants. Nucleic Acids Res *40*, 7123-7131.

Wu, J., Anczuków, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res *41*, 5149-5163.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859-1875.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., *et al.* (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J Neurosci *34*, 11929-11947.

Zhang, Z., Pinto, A.M., Wan, L., Wang, W., Berg, M.G., Oliva, I., Singh, L.N., Dengler, C., Wei, Z., and Dreyfuss, G. (2013). Dysregulation of synaptogenesis genes antecedes motor neuron pathology in spinal muscular atrophy. Proc Natl Acad Sci USA *110*, 19348-19353.

Zhu, L., Wang, X., Li, X.L., Towers, A., Cao, X., Wang, P., Bowman, R., Yang, H., Goldstein, J., Li, Y.J., *et al.* (2014). Epigenetic dysregulation of SHANK3 in brain tissues from individuals with autism spectrum disorders. Hum Mol Genet *23*, 1563-1578.

FIGURE LEGENDS

Figure 1 – An extensive program of neural-regulated AS

A) Distribution by type of human AS events with increased/decreased neural inclusion of the alternative sequence. Alt3/5, alternative splice site acceptor/donor selection; IR, intron retention; Microexons, 3-27 nt exons; Single/Multi AltEx, single/multiple cassette exons. B) Predicted impact of non-neural and neural-regulated AS events on proteomes. Neural-regulated events are more often predicted to generate isoforms preserving open reading frame (ORF) when the alternative sequence is included and excluded ("ORF-preserving isoforms", black), than to disrupt ORFs (i.e. the exon leads to a frame shift and/or introduces a premature termination

codon) specifically in neural samples ("ORF disruption in brain", dark grey) or in non-neural samples ("ORF preservation in brain", light grey). See Extended Experimental Procedures for details. C) Enrichment map for GO and KEGG categories in genes with neural-regulated AS that are predicted to generate alternative protein isoforms (top), and representative GO terms and their associated enrichment p-value for each subnetwork (bottom). The node size is proportional to the number of genes associated with the GO category, and the width of the edges to the number of genes shared between GO categories.

Figure 2 – A landscape of highly conserved neural microexons

A) Difference in exon inclusion level (ΔPSI) between the average PSIs for neural samples and non-neural samples (Y-axis) for bins of increasing exon lengths (X-axis). Microexons are defined as exons with lengths of 3-27 nt. Restricting the analysis to alternative exons with a PSI range across samples of >50 showed a similar pattern (data not shown). B) Number of exons by length whose inclusion level is higher (blue), lower (red) or not different (grey) in neural compared to non-neural samples. Short exons tend to be multiple of 3 nts and have higher inclusion in neural samples. C) Percent of neural-regulated microexons (of lengths of 3-15 and 16-27 nt) and longer exons that are predicted to generate alternative ORF-preserving isoforms (black), disrupt the ORF in/outside neural tissues (dark/light grey), or overlap non-coding sequences (white). D) Higher evolutionary conservation of alternative microexons compared to longer alternative exons at the genomic, transcriptomic (i.e. whether the exon is alternatively spliced in both species), and neural-regulatory level. Y-axis shows the percent of conservation at each specific level between human and mouse. p-values correspond to two-sided proportion tests. E) Percent of alternative microexons and longer exons that are detected as neural-regulated (average absolute ΔPSI>25) in each vertebrate species. F) Alternative 3-15 and 16-27 nt microexons show higher average phastCons scores at their intronic boundaries than longer alternative and constitutive exons. See also Figure S2.

Figure 3 – Switch-like regulation of microexons during neuronal differentiation

A) Heatmap of PSI changes (ΔPSIs) between time points during differentiation of ESCs to glutamatergic neurons *in vitro* (Hubbard et al., 2013). Yellow/pink indicate increased/decreased PSI at a given transition (T1 to T5). Unsupervised clustering detects eight clusters of exons based

on their dynamic PSI regulation (clusters I-VIII, legend). Right, top: scheme of the neuronal differentiation assay time points of sample collection, and analyzed transitions. Right, bottom: PSIs for each microexons (grey lines) in five selected clusters; red lines show the median for the cluster at each time point. B) Representative RT-PCR assays monitoring AS patterns of microexons during neuronal differentiation in Ap1s2 (9 nt), Mef2d (21 nt), Apbb1 (6 nt), Ap1b1 (21nt), Enah (12 nt) and Shank2 (9 and 21 nt). See also Figure S3.

Figure 4 – nSR100 is a positive, direct regulator of most microexons

A) Percent of neural-regulated exons within each length class that is affected by nSR100 expression in human 293T kidney cells (absolute ΔPSI > 15 [orange] or absolute ΔPSI > 25 [red]). p-values correspond to two-sided proportion tests of affected vs. non-affected events. B) Average normalized density of nSR100 cross-linked sites in 200 nt windows encompassing neural-regulated exons of different length classes. FPB, Fragments Per Billion. C) Cumulative distribution plots indicating the position of the first UGC motif within 200 nts upstream of neural-regulated microexons and longer exons, as well as non-neural and constitutive exons. $p<0.0001$ for all comparisons against microexons, Wilcoxon Sum Rank test. See also Figure S4.

Figure 5 – Microexons possess distinct protein-coding features

For each analysis, values are shown for neural-regulated, 3-15 nt microexons and longer (>27 nt) exons, as well as non-neural AS exons (see Figure S5 for other types of exons). A) Percent of exons with a high average (>0.67), mid-range (0.33 to 0.67) and low disorder rate (<0.33). B) Fraction of amino acids (AA) that overlap a PFAM protein domain. C) Percent of AA within PFAM domains predicted to be on the protein surface. D) Percent of AA types based on their properties; p-values correspond to the comparison of charged (acid and basic) versus uncharged (polar and apolar) AAs. E) Percent of exons that are adjacent to a domain (within 0-5 (black) or 6-10 AAs (grey)); p-values correspond to the comparison of exons within 0-5 AAs. F) Percent of residues overlapping PFAM domains involved in linear motif or lipid binding. G) Percent of residues overlapping binding motifs predicted by ANCHOR. H) Percent of exons with proteins identified as belonging to one or more protein complexes (data from (Havugimana et al., 2012)). All p-values correspond to proportion tests except for A (3-way Fisher test) and C (Wilcoxon Sum Ranks test). See also Figure S5.

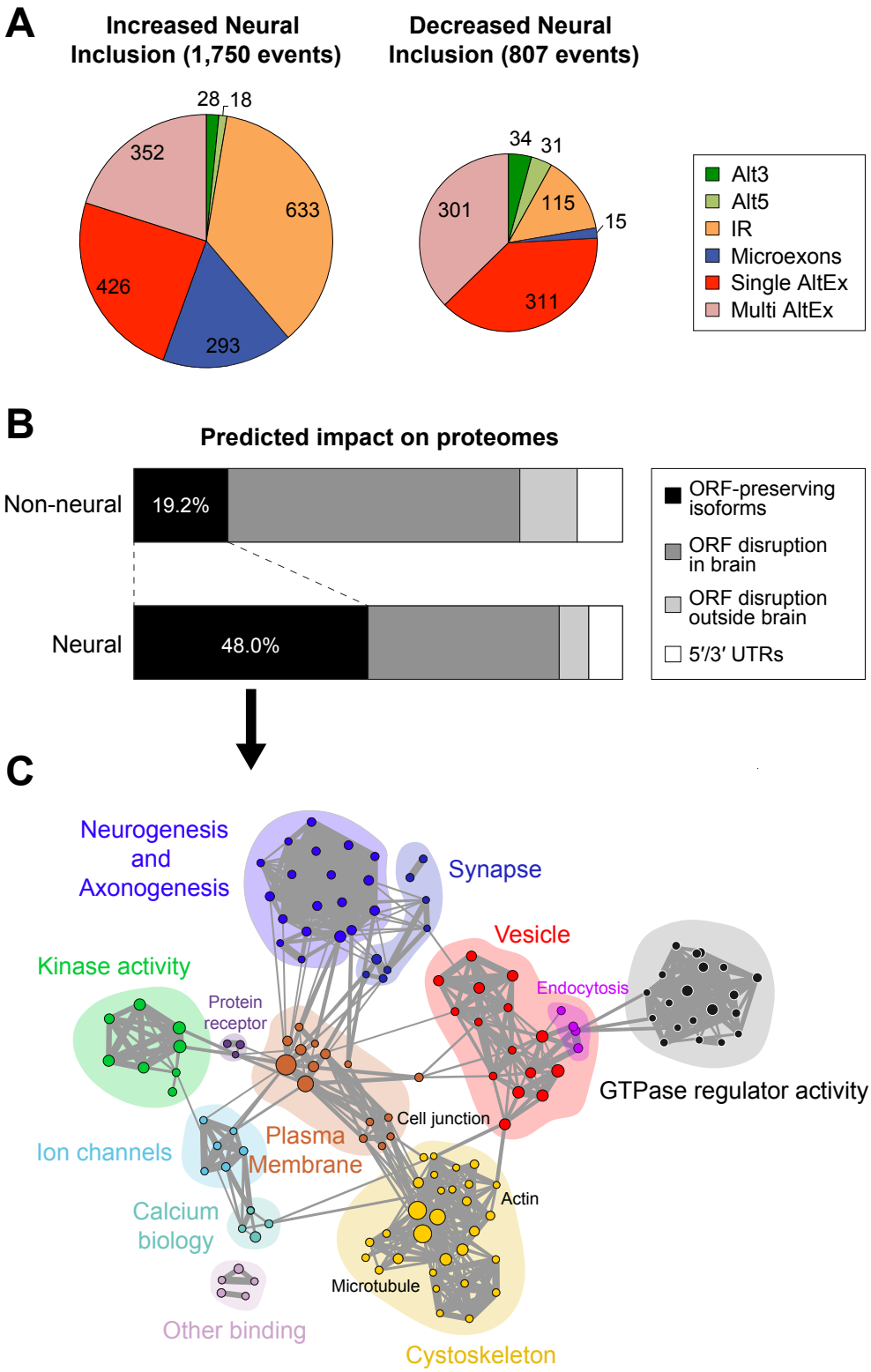
Figure 6 – Microexons regulate protein-protein interactions

A) Structural alignment of APBB1-PTB1 (pink) and APBB1-PTB2 (cyan) domains. Residues located at the protein-binding interface of APBB1-PTB2 are shown in blue. Inset shows the microexon residues in APBB1-PTB1 (E462-R463). B) Upon superimposition of APBB1-PTB1 (pink) and APBB1-PTB2 (cyan) domains, the microexon (magenta) is located close to the APBB1-PTB2 binding partner (APP protein fragment, blue), suggesting the microexon in PTB1 may affect protein binding. C) Quantification of LUMIER-normalized luciferase intensity ratio (NLIR) values for RL-tagged Apbb1, with or without the microexon, or with a mutated version consisting of two Alanine substitutions (ALA-mic.), co-immunoprecipitated with 3Flag-tagged Kat5. D, E) 293T cells were transfected HA-tagged Apbb1 (D) or AP1S2 (E) constructs, with or without the respective microexon, together with 3Flag-tagged Kat5 (D) or AP1B1 (E), as indicated. Immunoprecipitation was performed with anti-Flag (D) or anti-HA (E) antibody, and the immunoprecipitates were blotted with anti-HA or anti-Flag antibody, as indicated. Results shown in (E) were confirmed in a biological replicate experiment (Figure S6D). p-values in C and D correspond to t-tests for four and three replicates, respectively; error bars indicate standard error. Asterisk in panel E indicates a band corresponding to the light chain of the HA antibody.

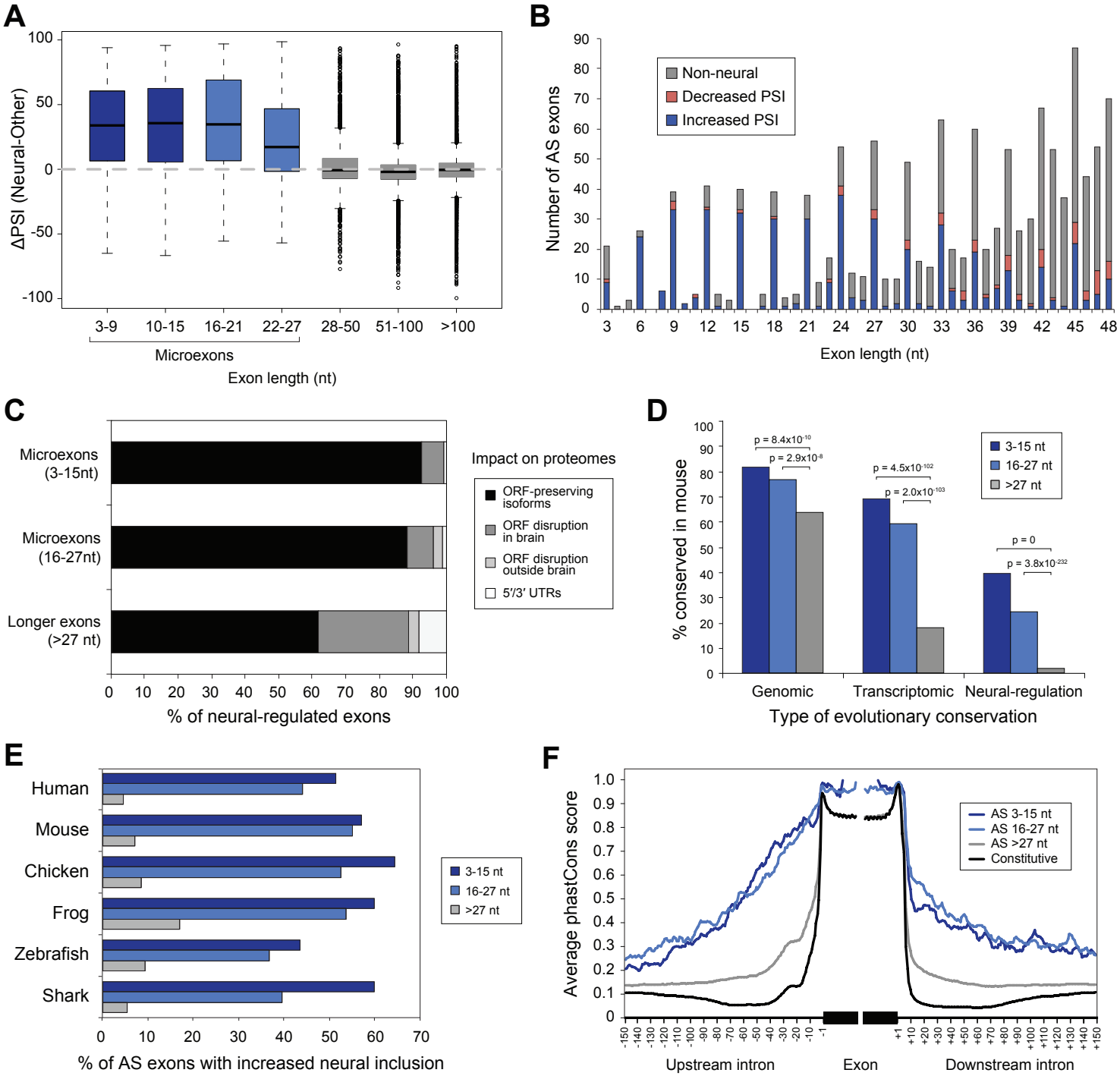
Figure 7 – Microexons are often misregulated in ASD

A) Percent of alternative exons of each length class that are misregulated in ASD (absolute ΔPSI>10 between PSI-averaged ASD and control groups. in ba41/42/22 brain regions. Dark shading, lower inclusion in ASD; light shading, higher inclusion in ASD; p-values correspond to proportion tests. B) Expression of nSR100 across the 12 control and 12 ASD individuals, Adjusted Fragments Per Kilobase Of Exon Per Million Fragments Mapped (FPKMs) were calculated using a regression analysis that accounts for variation derived from differences in RNA integrity, brain sample batch, sequencing depth, and 5' - 3' bias in measurements of gene-level FPKM values. C) Percent of exons within each length class misregulated in autistic compared to control brains (average absolute ΔPSI>10) for nSR100-regulated (ΔPSI>25 in the nSR100-overexpressing compared to control 293T cells) and non-nSR100-regulated (absolute ΔPSI<5) exons. D) Distribution of correlation coefficients between PSIs and nSR100 expression values across stratified ASD and control samples for microexons that are (n=59) or are not

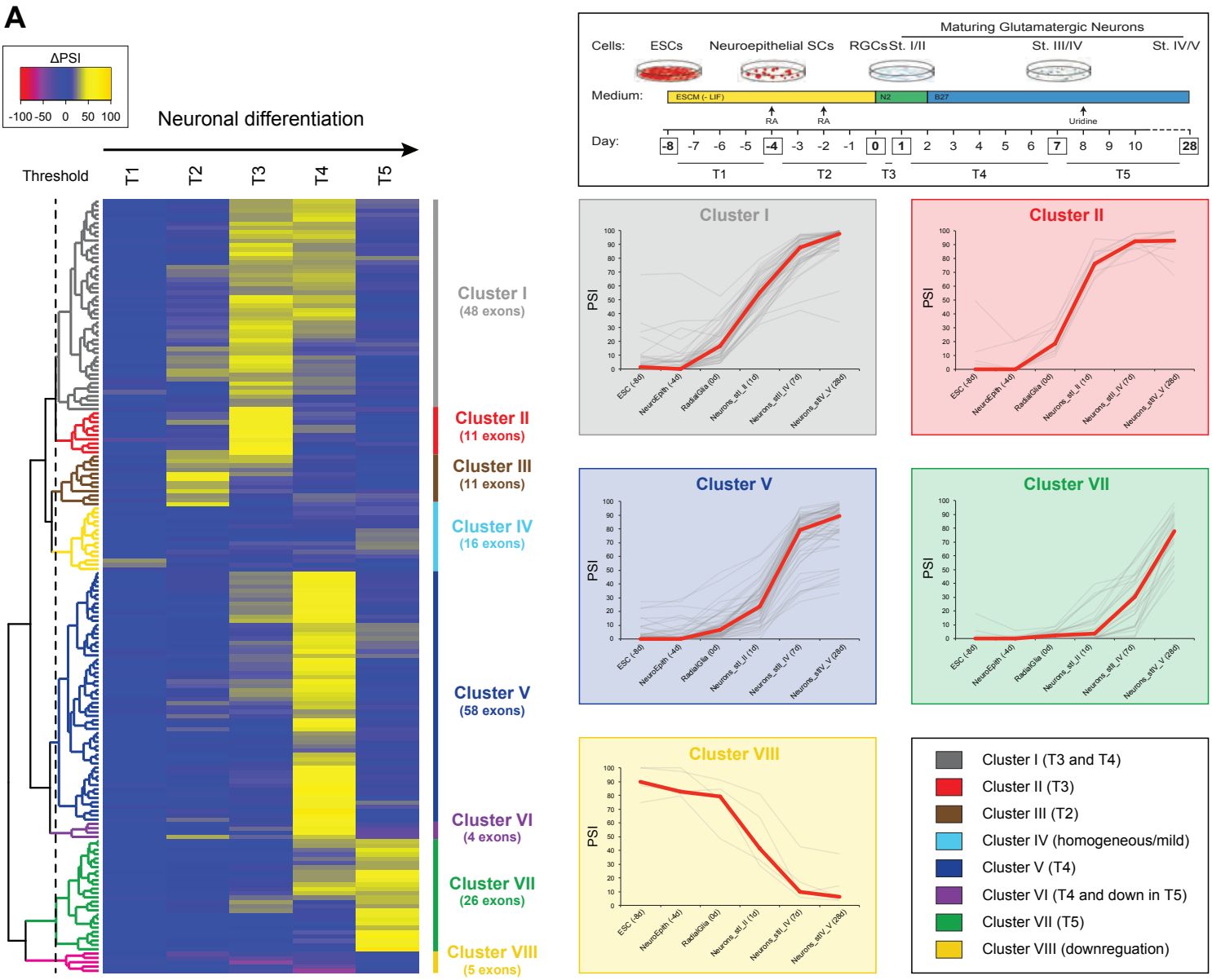
(n=69) regulated by nSR100. Only microexons with sufficient read coverage to derive accurate PSI quantifications in at least 9 ASD and 9 control ba41/42/22 samples were included. p-value correspond to Wilcoxon Sum Rank test. E) GO categories significantly enriched in genes with microexons that are misregulated in ASD. F) A protein-protein interaction network involving genes with ASD misregulated microexons (ΔPSI > 10) in ba41/42/22 brain regions. Genes with major effect mutations, and smaller effect risk genes, are indicated in red and shaded ovals, respectively. Genes grouped by functional category are indicated. See also Figure S7.

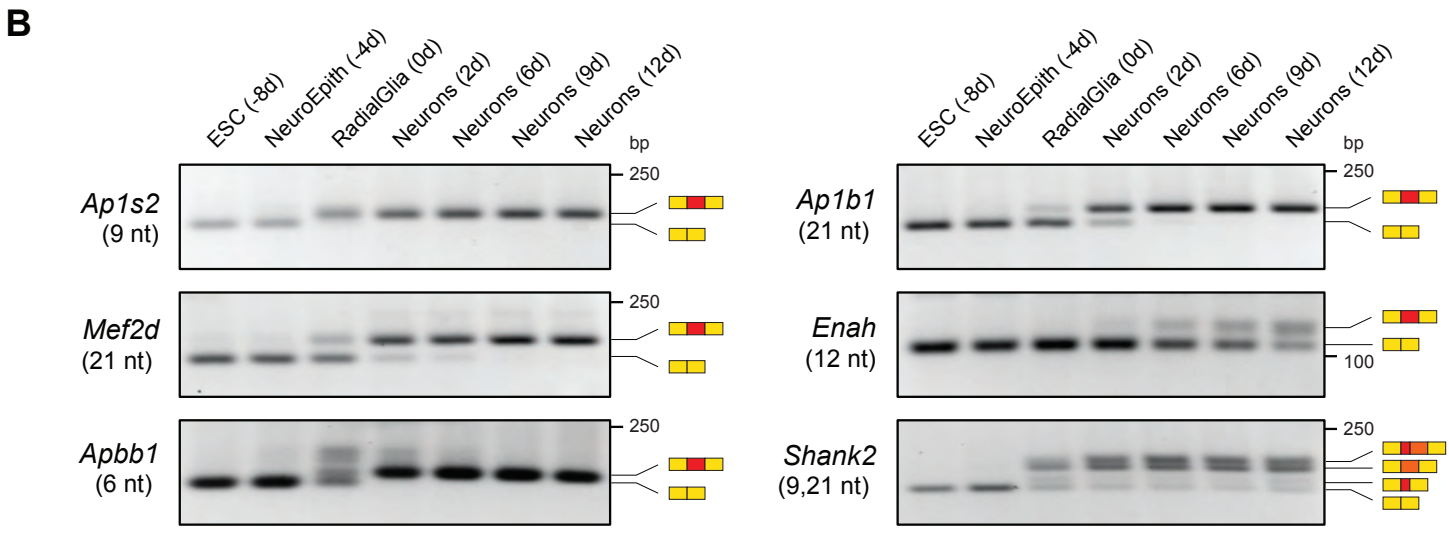


Irimia *et al.* Figure 1

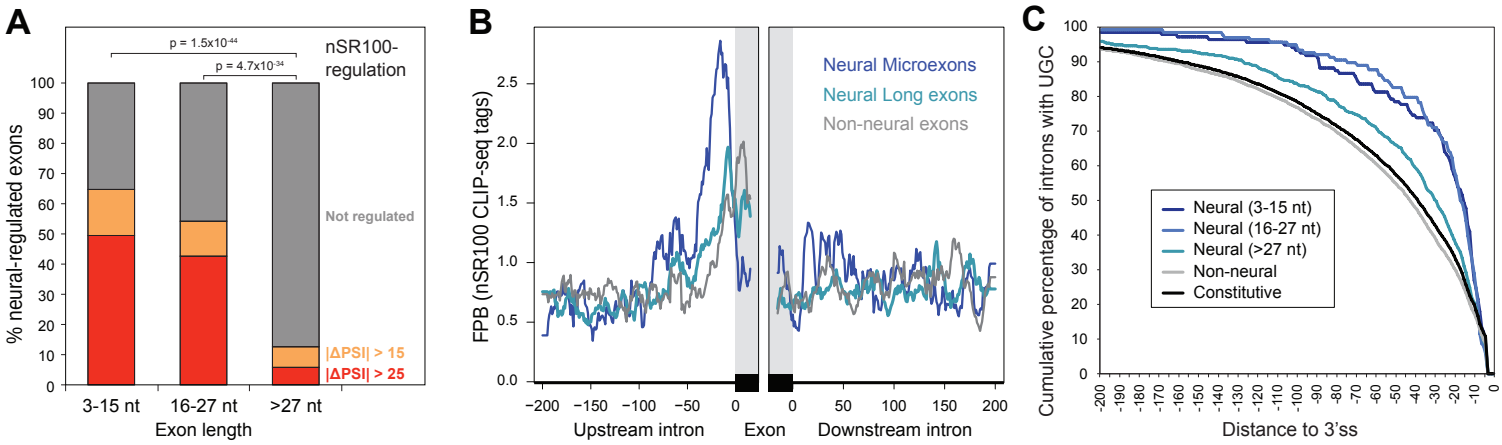


Irimia et al, Figure 2

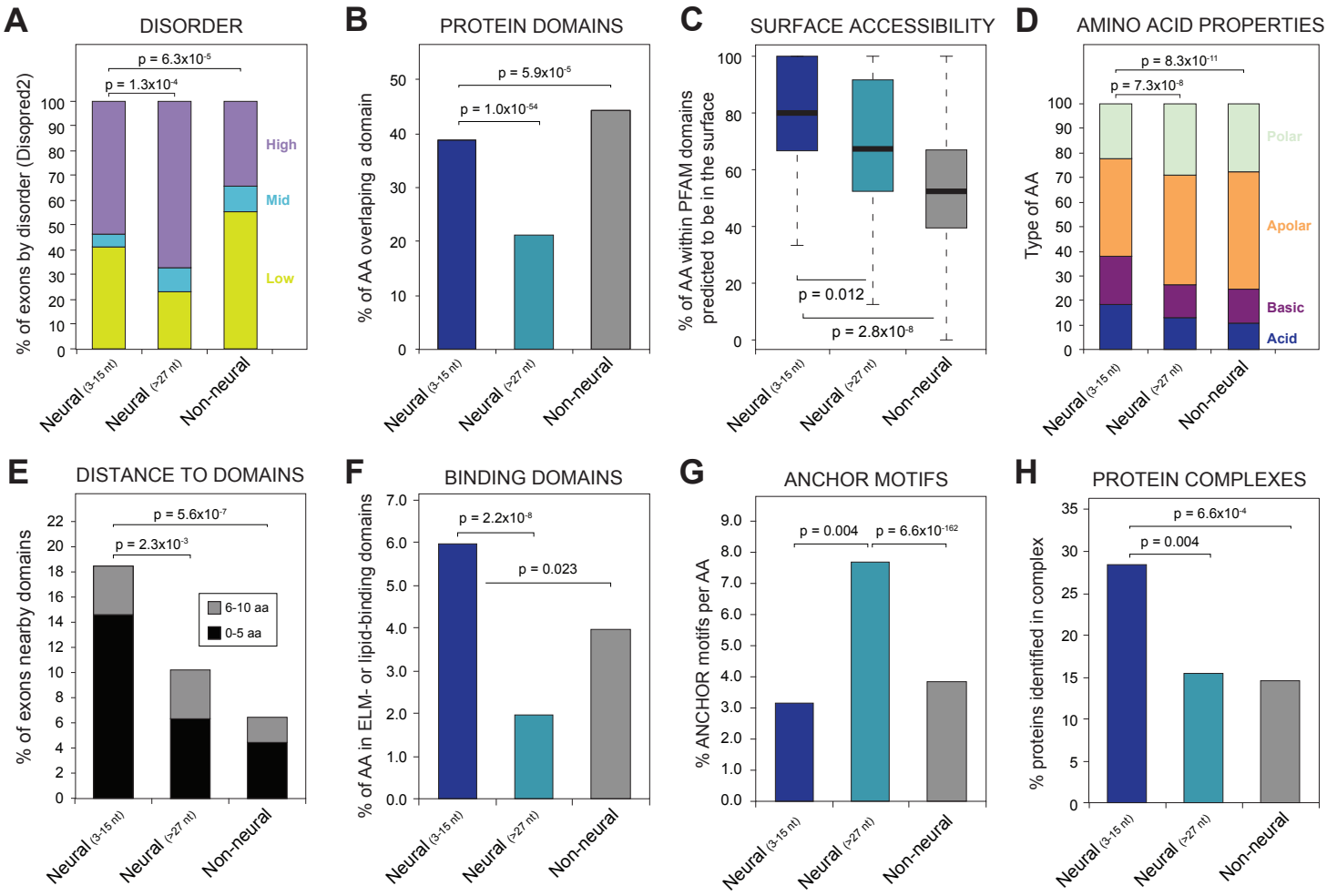




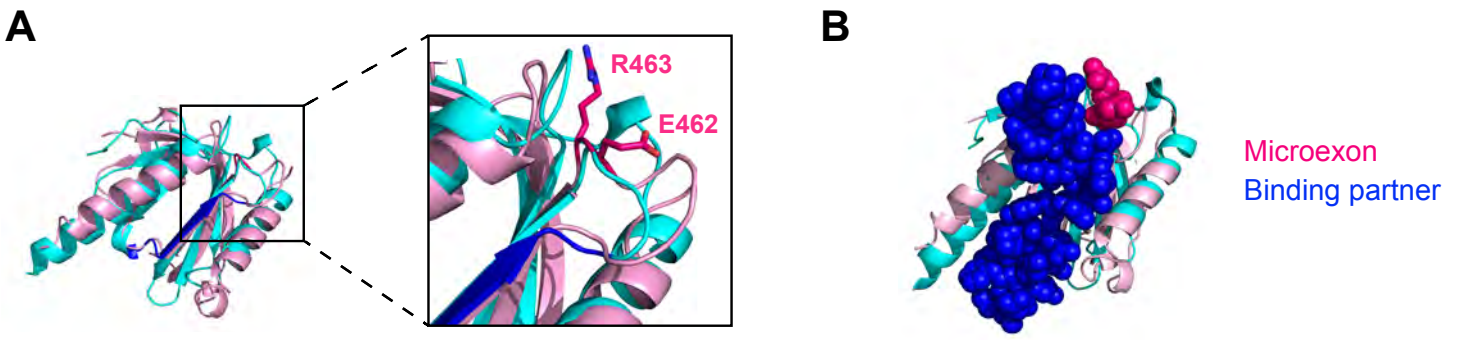
Irimia et al, Figure 3

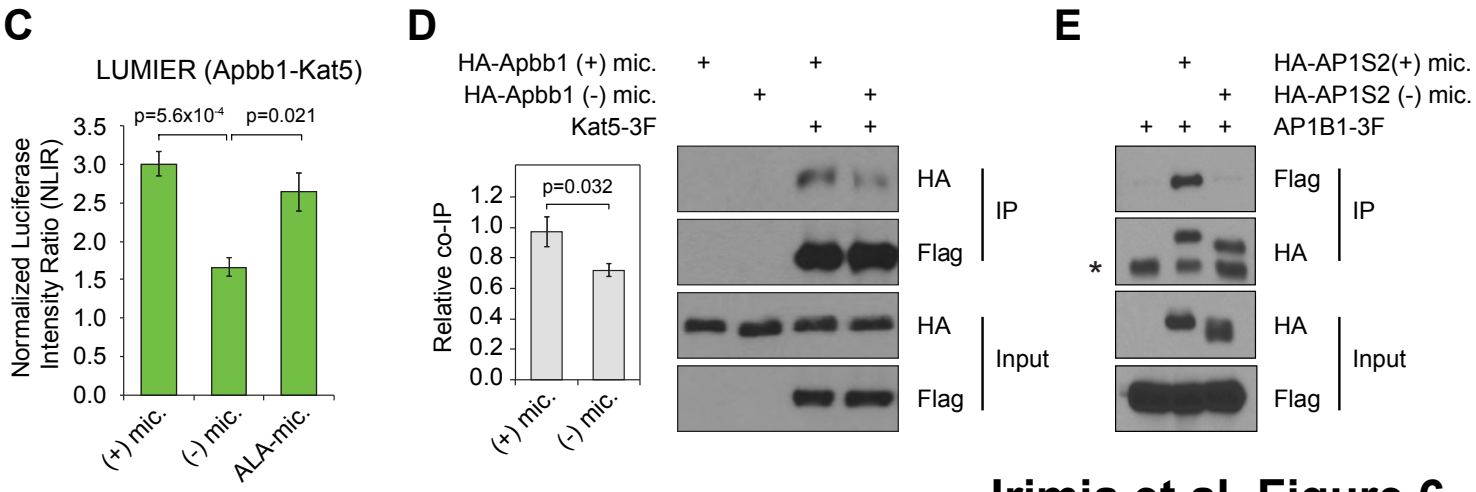


Irimia et al, Figure 4

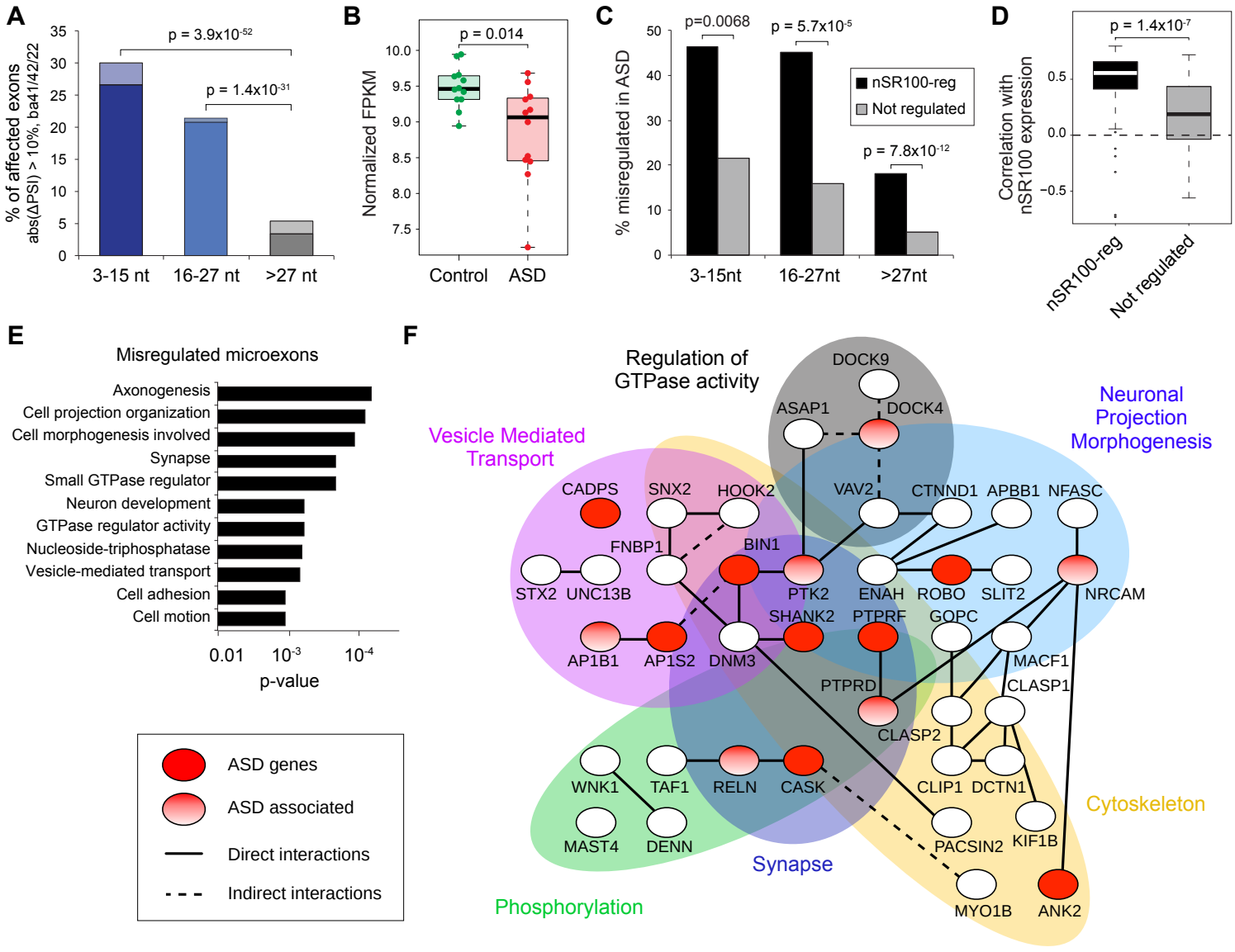


Irimia et al, Figure 5





Irimia et al, Figure 6



Irimia et al, Figure 7

Supplemental Information

A highly conserved program of neuronal microexons is misregulated in autistic brains

Manuel Irimia, Robert J. Weatheritt, Jonathan Ellis, Neelroop N. Parikshak, Thomas Gonatopoulos-Pournatzis, Mariana Babor, Mathieu Quesnel-Vallières, Javier Tapial, Bushra Raj, Dave O'Hanlon, Miriam Barrios-Rodiles, Michael J.E. Sternberg, Sabine P. Cordes, Frederick P. Roth, Jeffrey L. Wrana, Daniel H. Geschwind, and Benjamin J. Blencowe

SUPPLEMENTARY FIGURE LEGENDS

Figure S1 – Relationship between neural regulation at the AS and gene expression (GE) levels. Relates to Figure 1.

A) Overlap between differentially regulated genes at each level of regulation (GE and AS). Only 8.5% of the genes undergoing neural-regulated AS also display neural regulation at the GE level. B) Overlap of significantly enriched GO terms (Benjamini corrected p-value < 0.01) for genes that are significantly differentially upregulated at the mRNA steady-state levels in neural samples ("GE upregulated") and genes that harbor AS events that are differentially regulated in neural vs. non-neural samples and are predicted to generate alternative protein isoforms ("Alt. protein isoforms"). Over 40% of the GO categories enriched among the genes with neural-regulated AS are shared with those of genes upregulated at the GE level in neural tissues. p-values correspond to hypergeometric tests.

Figure S2 – Impact on protein and evolutionary conservation of neural-regulated exons. Relates to Figure 2.

A) Representative RT-PCR assays monitoring AS patterns of microexons in *Vav2, Rapgef6, Itsn1, Rims2, Abi1, Ptprd, Nbea, Zmynd8, Ppfia2 and Dnm2* (non-neural) in mouse neural (hippocampus, cerebellum and spinal cord), muscular (heart and skeletal muscle) and other (stomach, liver, spleen, kidney and testis) tissues. Molecular weight markers are indicated. B) For each sample, percent of microexons that show inclusion levels similar to neural (blue) or non-neural (red) samples. C) PSI distributions for neural-regulated microexons with increased neural

inclusion for different classes of cell and tissue types. For clarity, outliers are not shown. D) Heatmap of PSI changes (ΔPSIs) between time points during differentiation of C2C12 myoblasts to myotubes *in vitro* (Trapnell et al., 2010). Yellow/pink indicate increased/decreased PSI at a given transition (T1 to T3). Unsupervised clustering detects a cluster of 17 microexons with increased PSI during differentiation, particularly at T1. Right inset: PSIs for each microexons (grey lines) in the highlighted cluster; red line shows the median PSI at each time point E) Higher evolutionary conservation of neural 3-15-nt (dark blue) and 16-27-nt (lighter blue) microexons compared to longer neural exons (light blue) at the genomic, transcriptomic and neural regulatory level. Y-axis shows the percent of conservation between human and mouse. p-values correspond to proportion tests. F) Contribution of each type of AS events to events with conserved neural regulation between human and mouse, according to their predicted impact on proteomes. Microexons comprise approximately one third of all conserved neural events predicted to generate alternative protein isoforms. G) Distributions of average phastCons scores for exonic sequences of alternative microexons and long exons, as well as constitutive exons. H) Distributions of average phastCons scores for exonic sequences of neural-regulated microexons and long exons, as well as non-neural alternative exons and constitutive exons. p-values for G and H correspond to Wilcoxon Rank Sum tests. I) Average phastCons scores for neighboring intronic sequences of neural-regulated microexons and long exons, as well as non-neural alternative exons and constitutive exons. Only exons conserved at the genomic level between human and mouse were used for this analysis.

Figure S3 – Switch-like regulation of microexons during neuronal differentiation. Relates to Figure 3.

A) Heatmap showing relative gene expression level for key ESC and neural markers, including proneural genes (Neurog2 to Pax6) and post-mitotic neuronal markers (Elavl3/HuC and Rbfox3/NeuN). B) Distribution of relative ΔPSI (ΔPSI divided by the total PSI range across the six time points) for neural microexons and longer exons at each transition.

Figure S4 – Regulation of neural exons and microexons by splicing factors. Relates to Figure 4.

Percent of neural-regulated exons within each length class that is affected at 15< |ΔPSI| <25 (orange) and |ΔPSI| > 25 (red) by (A) RBFOX1 knock down in human neural precursor cells; (B) MBNL1 and MBNL2 double knock down in human HeLa cells; (C) ESRP1 knock down in human PNT2 cells; (D) nSR100/Srrm4 knock down in mouse N2A cells; (E) Ptbp1 knock down in mouse N2A cells; (F) Ptbp2 knock out in mouse cortex (P1 stage); (G) Ptbp2 knock out in mouse embryonic brain (18.5 days post conception); and (H) Rbfox1 knock out. p-values correspond to two-sided proportion tests of regulated vs non-regulated events. I) Expression of nSR100 in different isolated brain cell types (Zhang et al., 2014). J) Box plots comparing the 3′ and 5′ splice site strengths of neural 3-15 nt (dark blue) and 16-27 nt (light blue) microexons, longer (>27 nt, cyan) exons, non-neural alternative exons (grey), and constitutive exons (black).

Figure S5 – Protein features of different exon classes. Relates to Figure 5.

For each analysis, values are shown for neural 3-15 nt (dark blue) and 16-27 nt (light blue) microexons and longer (>27 nt, cyan) exons, as well as non-neural AS exons (grey) and constitutive exons (black). A, B) Percent of exons with high (average disorder rate >0.67), mid (between 0.33 and 0.67) and low (<0.33) disorder calculated using Disopred2 (A) and IUPred (B); p-values correspond to 3-way Fisher tests. C) Average disorder rate calculated using Disopred2 for each group of exons, as well as their neighboring upstream (C1, left) and downstream (C2, right) exons. D) Distribution of disorder rate across exon groups, calculated by IUPred. E) Percent of residues that overlap a PFAM protein domain. p-values correspond to proportion tests. F) Percent of AA within PFAM domains predicted to be in the protein surface using NetSurfP; p-values correspond to Wilcoxon Sum Ranks test. G) Accessible surface area score, based on the subset of exons with available crystal structures in PDB; p-values correspond to Wilcoxon Sum Ranks test. H) Percent of AA groups based on their properties; p-values correspond to proportion tests for the comparison of charged (acid and basic) versus uncharged (polar and apolar) AAs. I) Significantly enriched (Glu, Lys, Arg) or depleted (Pro, Thr) AAs in microexons compared to other exon types. Asterisks correspond to different levels of statistical significance (*, p<0.05; **, p<0.01; ***, p<0.001) in a proportion test. J) Percent of exons that fall nearby PFAM protein domains, without overlap. Black, within 0-5 AAs; grey, within 6-10 AAs. p-values correspond to proportion tests for exons within 0-5 AAs of a domain. K) Cumulative distance of exons that do not overlap domains with the nearest protein domain.

Exons in proteins with no predicted PFAM domain are excluded. L) Percent of residues overlapping PFAM domains involved in linear motif or lipid binding (Extended Experimental Procedures); p-values correspond to proportion tests.M) PFAM protein domains enriched in genes containing microexons and miniexons. Color scheme: red, protein binding GO (GO:0005515); dark pink, specific protein binding GO terms; orange, other binding GO terms; black, other or no associated GO terms. Y-axis corresponds to DAVID p-values. N) Percent of residues overlapping ANCHOR binding motifs; p-values correspond to proportion tests. O) Degree (number of interactors in PPI networks) of proteins containing different types of exons. Degree values obtained from (Ellis et al., 2012). p-values correspond to Wilcoxon Sum Ranks test. P) Percent of exons in which the containing proteins have been identified as part of protein complexes (data from (Havugimana et al., 2012)); p-values correspond to proportion tests.

Figure S6 – Location of microexons in protein structures. Relates to Figure 6.

A) Selection of available protein structures from PDB and SWISS-MODEL, and modeled structures using Phyre containing neural-regulated microexons (in red). The number of residues of each microexon is indicated in parenthesis. B, D) 293T cells were transfected HA-tagged Apbb1 (B) or AP1S2 (D) constructs, with or without the microexon, together with 3Flag-tagged App (B) or AP1B1 (D), as indicated. Immunoprecipitation was performed with anti-Flag antibody or anti-HA antibody, as indicated. C) Quantification of LUMIER-normalized luciferase intensity ratio (NLIR) values for RL-tagged Apbb1, with or without the microexon, co-immunoprecipitated with 3Flag-tagged App. p-values in B and C correspond to t-tests for three replicates, respectively; error bars indicate standard error.

Figure S7 – Microexons are often misregulated in ASD. Relates to Figure 7.

A) Percent of neural-regulated exons by length groups that are misregulated in ASD ($|\Delta PSI| > 10$ between averaged ASD and control groups) in ba41/42/22 brain region. p-values correspond to proportion tests. B) Representative RT-PCRs for microexons misregulated in ba41/42/22 and ba9 regions from ASD versus control individuals (see Table S1). Bottom: boxplot of isoform quantifications from RT-PCR assays for 10 microexons in control (n=70 data points) and ASD (n=80 data points) individuals. p-value from Wilcoxon Sum Ranks test. C) Heatmap and unsupervised clustering of z-scores of PSIs for microexons misregulated in ASD individuals with

sufficient read coverage in at least 9 ASD and 9 control ba41/42/22 samples (n=64), and of nSR100 expression values. Conditions: ASD (red), control (green). Asterisks indicate individual samples used in RT-PCR and qRT-PCR analyses (panels B and D). D) qRT-PCR quantifications of nSR100 expression in four ASD and three control ba41/42/22 samples (see panel C) normalized for three different housekeeping genes. p-values correspond to two-sided t-tests. E) PSI distributions of the 6-nt and 9-nt microexons in *APBB1* and *AP1S2*, respectively, in control (green) and ASD (red) individuals; p-values from Wilcoxon Sum Ranks test.

SUPPLEMENTARY TABLE LEGENDS

Table S1 – RNA-Seq samples used in this study

Table S2 – Neural-regulated AS events in human and mouse

Full coordinate: chromosome: C1 donor, AS exon, C2 acceptor. Alt3/5, alternative splice site acceptor/donor selection; IR, intron retention; AltEx, cassette alternative exons (including microexons; length $\leq$ 27 nt).

Table S3 – Highly conserved neural-regulated microexons across vertebrate species

For each species in which the orthologous microexon can be identified, it shows the average PSI in neural ('av_NEURAL') and non-neural ('av_REST') samples, as well as the sequence of the microexon (A), and constitutive upstream (C1) and dowsntream (C2) exons.

Table S4 – Gene Ontology analyses for clusters of microexons based on their PSI dynamics during neuronal differentiation

Table S5 – Features of human alternative microexons

In "Basic Protein Information", EEJ=# in start and stop positions represents insertion point of exon that could not be mapped to Ensembl protein (i.e. position of the exon-exon junction between C1 and C2 exons.

In "Protein features (various)", the annotation with "NA-START-END" represents the annotation of C1 and C2 exons combined (i.e. with the A exon removed). Empty cell indicates that the A exon could be mapped, but no specific annotation for the feature was present.

In "Disorder Predictions", predictions are provided for Disopred2 and IUPred

In "Protein Domain Information", the UniProt annotation starts with assigned position within the canonical UniProt protein (either length of A exon or where exon is inserted) and afterwards is comma separated with assignments from UniProt Annotation. In addition to UniProt and Pfam annotations, *de novo* Pfam and PROSITE predictions are provided for C1, A and C2 exons. Columns "Dist_5" and "Dist_3" indicate the distance to the closest domain upstream and downstream, respectively (OV, overlapping a domain). References: Pfam (Finn et al., 2014); PhosphoSite (Hornbeck et al., 2004); Transmembrane predictions, TMHMM (Krogh et al., 2001); ELM (Dinkel et al., 2014); SignalP (Petersen et al., 2011); COIL (Lupas et al., 1991); Disorder, IUPred (Dosztányi et al., 2005), Disopred2 (Ward et al., 2004); CAST (Promponas et al., 2000); ANCHOR (Dosztányi et al., 2009); UniProt (Consortium., 2014).

Table S6 – Inclusion levels for human microexons across 52 cell and tissues types
For each sample, two columns are provided: percent inclusion (PSI) and coverage (in order of increasing read coverage, N<VLOW<LOW<OK<SOK; VLOW corresponds to the "minimum coverage" criteria specified in Extended Experimental Procedures).

EXTENDED EXPERIMENTAL PROCEDURES

RNA-Seq data and genomes

Unless stated otherwise, RNA-Seq data was from Poly(A)$^+$ RNA. A full list of analyzed RNA-Seq datasets is in Table S1. Analyses used the following genome releases: *Homo sapiens*, hg19, *Mus musculus*, mm9; *Gallus gallus*, galGal3; *Xenopus tropicalis*, xenTro3; *Danio rerio*, danRer7; *Callorhinchus milii*, v1.0 (Venkatesh et al., 2014)) (Table S1).

To increase the fraction of reads mapping within each RNA-Seq sample, each read was first split into 50-nucleotide (nt) read groups. If the original reads were ≥100 nt, two 50-nt reads were generated by trimming from the 3' and 5', respectively; if reads were >50 nt but <100 nt, two overlapping 50-nt reads were produced. In addition, in case of paired-end sequencing, both read mates were pooled. These 50-nt split reads were then mapped to the corresponding genome assemblies using Bowtie (Langmead et al., 2009), using –m 1 –v 2 parameters (unique mapping with no more than two mismatches). Reads that mapped to the genome were discarded for alternative splicing (AS) quantifications based on exon-exon junction (EEJ) counts (see below). For quantification, only one count per read group (i.e. all sub-reads coming from the same original read) was considered to avoid multiple counting of the same original read.

Alternative splicing analysis pipeline

To comprehensively detect and quantify all major types of AS events involving alternative exons (AltEx), alternative 5´ and 3´ splice site selections (Alt5/Alt3), and intron retention (IR), we developed a multi-module analysis pipeline, which will be described in greater detail elsewhere (Irimia and Blencowe, 2014) and is publically available at https://github.com/vastgroup/vast-tools. This pipeline initially uses a wide range of inputs (RNA-Seq data, EST and cDNA evidence, gene annotations, and evolutionary conservation) to assemble libraries of EEJs for subsequent use in the detection and quantification of alternatively spliced sequences from RNA-Seq reads. For AltEx, to capture single exon skipping events, as well as more complex events involving tandem exons with correlated, independent or mutually exclusive AS, three complementary modules were developed for assembling libraries of EEJs: (i) A “transcript-based module”, employing alignments of ESTs, cDNAs and RNA-Seq-based transcript sequences (generated using cufflinks (Trapnell et al., 2010)) with genomic sequence to identify transcript-supported EEJs (essentially as described in (Khare et al., 2012)); (ii) A “splice site-based module”, utilizing the joining of all hypothetically-possible EEJ combinations from annotated and *de novo* splice sites (essentially as described in (Han et al., 2013), with a few species-specific modifications to incorporate *de novo* splice sites, see below); (3) A “microexon module”, employing *de novo* searching of pairs of donor and acceptor splice sites in intronic sequence to detect novel, very short (i.e. 3-15 nt) microexons (see details below). Each of these modules was

used to generate an independent library of EEJs to derive measurements for exon inclusion levels using the metric "Percent Spliced In" (PSI). The results are then combined to generate a unique, non-redundant set of PSI values for all AltEx events in a given species. To detect and quantify Alt3 and Alt5 events, we used the output from mapping RNA-Seq reads to the EEJ library generated by the "splice site-based module", which provides information on the usage of every hypothetical splice site donor and acceptor. For each alternative splice site in an Alt3 or Alt5 event (defined as two or more tandem splice site acceptors or donors, respectively), we calculated the "Percent Splice-site Usage" (PSU) based on the fraction of reads supporting the usage of each alternative splice site. Finally, for IR, we used our recently described pipeline (Braunschweig et al., 2014), which employs a comprehensive set of reference sequences comprising exon-intron junctions (EIJs), intron mid-point sequences, and EEJs formed by intron removal. Introns were classified as retained when there was a balanced accumulation of reads mapping to 5´ and 3´ EIJs and the mid-intron sequence, relative to the EEJ sequence, and the level of retention was represented using the metric "Percent Intron Retention" (PIR). For all modules and AS types, quantifications were done based on read counts corrected for the number mappable positions in each EEJ or EIJ following the formula:

$$Corrected_EEJ_{count} = EEJ_{count} * \frac{Maximum_{mappability}}{EEJ_{mappability}}$$

where EEJ_{count} is the number of read groups mapped to the EEJ, $Maximum_{mappability}$ the maximum number of mapping positions that any EEJ can have for reads of length 50 nt, and $EEJ_{mappability}$ the number of positions that can be mapped uniquely to the EEJ using specific bowtie parameters (–m 1 –v 2), and thus $EEJ_{mappability} \leq Maximum_{mappability}$ (see (Barbosa-Morais et al., 2012; Han et al., 2013) for details).

To incorporate novel (i.e. not annotated) splice sites into the library of potential EEJs in the "splice site-based module", we performed the following steps. First, for human, we incorporated novel EEJs identified by ENCODE (Gene Expression Omnibus [GEO] Accession GSE30567). For these data, we restricted novel EEJs to acceptor-donor pairs separated by <300nt (i.e. a putative novel exon of length <300 nt), and joined to previously annotated splice sites. For other

species, for which ENCODE-like data are not available, we performed a *de novo* search for splice sites using our RNA-Seq dataset (Table S1), as previously described (Curtis et al., 2012). In short, for each annotated splice site donor/acceptor, we scanned two downstream/upstream introns for potential splice site acceptors/donors that would constitute a novel EEJ supported by at least five reads, mapped to multiple positions of the EEJ. Finally, for all species but shark, we also performed pairwise lift-overs using Galaxy to obtain those EEJs conserved at the genomic level (see "Evolutionary analysis of neural-regulated cassette exons" below) that are annotated in some species but not in others (e.g. annotated in human but not in chicken). Lift-overs could not be performed from/to shark due to lack of genomic-wide alignments. This enriched set of splice sites was then used to produce a library, for each species, containing all possible forward combinations of donor-acceptors within each gene, and detection and quantification of cassette exons using corrected read counts was done as previously described (Han et al., 2013).

- Microexon module: since in our EEJ quantifications we require that a minimum of eight positions from each exon that is part of the EEJ overlaps any mapping read (Han et al., 2013), the inclusion of very short exons (<8 nt) cannot, by definition, be quantified. Moreover, exons shorter than 15 nt have very few mappable positions (8 or fewer) over their inclusion-supporting EEJs, often resulting in unreliable quantifications. Therefore, to investigate the inclusion of exons of between 3 and 15 nts, we developed a “microexon module”. This module employs similar criteria as the other modules to quantify transcripts with exon skipping; however, it uses exon-microexon-exon junction (EEEJ) read counts as a proxy for exon inclusion. These reads simultaneously map to the upstream (C1), alternative (A) and downstream (C2) exons, providing additional mappable positions supporting exon inclusion and thus a more accurate quantification of PSI. To generate such libraries of EEJs and EEEJs for microexons for each species, we first built a comprehensive microexon annotation. For this purpose, we scanned the intronic sequence between each pair of annotated exons in each canonical transcript (as defined by BioMart) for any potential 3-15-nt microexon (i.e. 3 to 15 nt flanked by AG and GT. Since searches for candidate 1 and 2-nt exons using similar procedures yielded a high rate of false-positives, largely due to genomic sequence variation and sequencing errors, we restricted our analyses to exons of at least 3 nt). Multi-fasta libraries of EEEJ candidates (between 17 and 46 million for each species) were built, and RNA-Seq data from every sample was mapped using Bowtie with highly

stringent conditions (-m 1 –v 0, uniquely mapping with no mismatches) to identify those EEEJs supported by at least one read. After filtering sequences that could be ascribed to Alt5 or Alt3 events (based on the genomic sequence adjacent to C1 and C2 exons, respectively) or that overlapped longer annotated exons, the putative microexons were combined with annotated microexons from Ensembl, and cross-species, pair-wise lift-overs were also performed to identified further genome-conserved microexons (except for shark). From this comprehensive set of microexons we then derived multi-fasta libraries containing all EEJs and EEEJs, calculated their mappability as described above, mapped RNA-Seq (using bowtie with -m 1 –v 2 parameters), and quantified microexon inclusion following the formula:

$$PSI\ (A) = 100 * \frac{\Sigma\,(\mathrm{C}_i A_{ij} \mathrm{C}_j)}{\Sigma\,(\mathrm{C}_i A_{ij} \mathrm{C}_j) + \Sigma\left(\mathrm{C}_i \mathrm{C}_j\right)}$$

where A_{ij} is a specific microexon variant (if associated with multiple alternative 5´ and/or 3´ splice sites, always producing a microexon), C_i is any possible splice donor upstream of the microexon, and C_j any possible splice acceptor downstream of the microexon; $C_iA_{ij}C_j$ is thus an EEEJ supporting microexon inclusion and C_iC_j an EEJ supporting microexons skipping.

For human and mouse, the outputs from the three AltEx modules were combined to produce a non-redundant list of cassette exons and associated quantifications. For exons that are identified by more than one module (~80%), the representative with the highest overall read coverage is kept. In case of equal coverage (most cases), priority is given to events derived from the "transcript-based" module, followed by those from the "microexon" module. For species other than human and mouse, we used only the "splice site-based" and "microexon" modules, due to the limited availability of full transcript data.

Alternative splicing definition and minimum read coverage

For all types of events, we used the same definition for a given sequence to be considered alternatively spliced: $10 \leq$ PSI/PSU/PIR ≤ 90 in at least 10% of the samples with sufficient read coverage and/or a range of PSI/PSU/PIRs ≥ 25 across all samples with sufficient read coverage.

A given event was considered to have sufficient read coverage in a particular RNA-Seq sample according to the following criteria:

- For AltEx (except for those quantified using the microexon pipeline): (i) ≥10 actual reads (i.e. before mappability correction) mapping to the sum of exclusion EEJs, ***OR*** (ii) ≥10 actual reads mapping to one of the two inclusion EEJs, and ≥5 to the other inclusion EEJ.

- For microexons: (i) ≥10 actual reads mapping to the sum of exclusion EEJs, ***OR*** (ii) ≥10 actual reads mapping to the sum of inclusion EEEJs.

- For IR: (i) ≥10 actual reads mapping to the sum of skipping EEJs, ***OR*** (ii) ≥10 actual reads mapping to one of the two inclusion EIJs, and ≥5 to the other inclusion EIJ.

- For Alt3 and Alt5: ≥10 actual reads mapping to the sum of all EEJs involved in the specific event.

Comparison with the Olego software

Wu et al. (Wu et al., 2013) have recently described Olego, an RNA-Seq splice mapper that uses short sequence seeds and that is able to identify exons that are as short as 9 nt. In their study, Wu et al. define microexons as exons of lengths ≤ 25 nt. Using an RNA-Seq sample from mouse retina, they report 613 alternative microexons (234 of length ≤ 15 nt), of which 334 (132) have PSI values between 10 and 90 (Wu et al., 2013). Using similar parameters, we detect 1,008 (523) microexons ≤ 25 nt, of which 868 (424) have 10≤PSI≤90 in at least one sample. Using only an RNA-Seq sample obtained from retina and eye tissues (see Table S1), we identify 424 (205) alternative microexons with 10≤PSI≤90. In total, 115 out of the 132 (87%) alternative microexons of length ≤ 15 nt reported by Wu et al (Wu et al., 2013) could be identified using our pipeline. In contrast, from the extensive collection of RNA-Seq data that we have profiled, we detect 408 microexons that were not detected by Wu and colleagues, including dozens of 3-6-nt microexons.

Definition of neural-regulated alternative splicing

To identify neural-regulated AS events, we first averaged PSI/PSU/PIR values of samples from similar tissues or cell types, as indicated in the column “Group” in Table S1. For an AS event to be considered, we required a minimum of 2 neural samples and 5 distinct non-neural samples/groups (referred below as “rest”) with sufficient read coverage (see above). We

implemented a set of non-mutually exclusive definitions to maximize the detection of different patterns of neural AS regulation, including both quantitative (i.e. with large PSI differences between the samples from the neural and non-neural groups) and qualitative (i.e. in which one of the isoforms is present only in the set of neural or non-neural samples, even if with relatively small PSI differences between both groups). These definitions were refined upon manual inspection of PSI profiles across samples in order to minimize false-positive calls. The following definitions were employed:

1) absolute ($Mean_{neural}$ - $Mean60_{rest}$) ≤ 25 ***AND***
   [$Range_{neural}$ < absolute ($Mean_{neural}$ - $Mean_{rest}$)/2 ***OR***
   Range_rest < absolute ($Mean_{neural}$-$Mean_{rest}$)/2 ***OR***
   SD_{neural} < absolute ($Mean_{neural}$ - $Mean_{rest}$)/2 ***OR***
   SD_{rest} < absolute ($Mean_{neural}$ - $Mean_{rest}$)/2]

2) $Total_samples_{rest}$ ≥ 10 ***AND***
   $Total_samples_{neural}$ ≥ 3 ***AND***
   ([$Mean_{neural}$ - $Mean60_{rest}$ ≤ -10 ***AND*** Max_{neural} < Min_{rest}] ***OR***
   [$Mean_{neural}$ - $Mean60_{rest}$ ≥ 10 ***AND*** Max_{rest} < Min_{neural}]) **AND**
   [SD_{neural} < absolute ($Mean_{neural}$ - $Mean_{rest}$)/2 ***OR***
   SD_{rest} < absolute ($Mean_{neural}$ - $Mean_{rest}$)/2]

3) absolute ($Mean_{neural}$ - $Mean80_{rest}$) ≤ 20 ***AND***
   [SD_{neural} < absolute ($Mean_{neural}$ - $Mean_{rest}$)/4 ***OR***
   SD_{rest} < absolute ($Mean_{neural}$ - $Mean_{rest}$)/4]

4) $Total_samples_{rest}$ ≥ 20 ***AND***
   $Total_samples_{neural}$ ≥ 3 ***AND***
   [(Min_{rest} ≥ 95 ***AND*** Max_{neural} < 90) ***OR***
   (Min_{neural} ≥ 95 ***AND*** Max_{rest} < 90)]

5) $Total_samples_{rest}$ ≥ 10 ***AND***

$Total_samples_{neural} \geq 3$ ***AND***

[($Mean_{neural}$ - $Mean80_{rest} \leq -10$ ***AND*** $Mean_{rest} > 98$) ***OR***

($Mean_{neural} - Mean80_{rest} \geq 10$ ***AND*** $Mean_{rest} < 2$)]

6) $Total_samples_{rest} \geq 10$ ***AND***

$Total_samples_{neural} \geq 3$ ***AND***

absolute ($Mean_{neural}$ - $Mean60_{rest}$) ≥ 15 ***AND***

$Range_{neural} < 10$

where $Total_samples_{neural/rest}$ is the number of neural/distinct non-neural samples with enough read coverage; $Mean_{neural/rest}$ is the mean PSI/PSU/PIR for all neural/distinct non-neural samples; $Mean60_{rest}$ and $Mean80_{rest}$ are the mean PSI/PSU/PIR for the distinct non-neural samples excluding the 40% or 20% of samples with the most distant PSI/PSU/PIRs from the neural mean value (i.e. 7^{th} to 10^{th} and 9^{th} to 10^{th} deciles), respectively; $Min_{neural/rest}$ is the minimum PSI/PSU/PIR value for neural/non-neural samples; $Max_{neural/rest}$ is the maximum PSI/PSU/PIR value for neural/non-neural samples; $Range_{neural/rest}$ is the difference between $Max_{neural/rest}$ and $Min_{neural/rest}$; and $SD_{neural/rest}$ is the standard deviation of PSI/PSU/PIR values for neural/distinct non-neural samples. In addition, we required bona fide neural-regulated AS events to have a $p < 0.05$ when comparing neural and non-neural distinct samples using the B-statistic, i.e. the empirical Bayes log-odds of differential PSI/PSU/PIRs (Smyth, 2004) (as implemented in "ebayes", from the limma package in R). "Non-neural" events were defined as those AS events (as defined in the section "Alternative splicing definition and minimum read coverage") that had (i) enough read coverage in at least 2 neural samples and 5 distinct non-neural samples, (ii) $|(Mean_{neural} - Mean_{rest})| < 10$, and (iii) $p \geq 0.05$ on the Bayesian test.

For the analysis across vertebrate species (Figure 2E), and to compensate for the different availability in RNA-Seq samples, we only required an $|\Delta PSI| \geq 25$ between the averaged neural and distinct non-neural samples for exons expressed in at least one neural sample and three distinct non-neural samples (Han et al., 2013). Using these conditions, we identified 543 alternative microexons in human, 530 in mouse, 430 in chicken, 209 in Xenopus, 230 in zebrafish, and 161 in elephant shark. However, it should be noted that read coverage and

diversity of available neural tissues varied widely across species, and this variation was found to correlate with total microexon numbers (but not with the fraction of those that are neural-regulated). Repeating our analyses by selecting eight equivalent tissue samples from human, mouse, chicken and zebrafish ("Subset" in Table S1) with similar read coverage yielded comparable numbers of alternative microexons across all species (human, 110; mouse, 198; chicken, 143; zebrafish, 136).

Neural-differentially expressed genes were obtained from (Braunschweig et al., 2014), based on Bayesian statistics of gene expression levels across the same panel of cell and tissue RNA-Seq samples.

Analysis of microexon inclusion and nSR100 expression in glial cells and neurons

To investigate the cell type specificity of neural microexon regulation, we used RNA-Seq data from isolated glial cells and neurons (Table S1, (Sofueva et al., 2013; Zhang et al., 2014; Zhang et al., 2013)). For each sample, we compared the PSI of each neural-regulated microexon with sufficient read coverage to derive a confident PSI value in that sample (PSI_{sample}) against its averaged neural PSI ($Mean_{neural}$, see above). We then scored the fraction of microexons that showed a PSI_{sample} similar to neural samples ($Mean_{neural}$ - PSI_{sample} < 10; blue in Figure S2B) or to non-neural samples ($Mean_{neural}$ - PSI_{sample} > 25; red in Figure S2B). For Figure S4I, nSR100 expression in different isolated neural cell types (Zhang et al., 2014) was calculated using the cRPKM metric (Labbe et al., 2012).

Gene Ontology analyses

The Gene Ontology (GO) network for the human genes harboring neural-regulated AS events predicted to generate alternative protein isoforms (Figure 1C) was built using Enrichment Map (Merico et al., 2010) in Cytoscape 2.8.3 (Shannon et al., 2003) with the following parameters: p-value cut-off = 0.001; FDR Q-value cut-off = 0.1; Jaccard+Overlap Combined option, with cut-off = 0.375; Combined Constant = 0.5. Enriched functional terms were obtained from DAVID chart for GO_FAT and KEGG pathways (http://david.abcc.ncifcrf.gov/)(Huang da et al., 2009), using multiexonic genes expressed in at least two neural samples and five distinct non-neural

samples as background. Clusters of functionally related terms were manually circled and assigned a label based on the representative GO term with the lowest p-value within the cluster.

To analyze functional categories enriched in microexons misregulated in autism spectrum disorder (ASD; Figure 7E), we used DAVID chart with the same options as above. The functional network of genes with microexons misregulated in ASD (Figure 7F) was built manually. The functional categories are based on annotation from the UniProt database and/or GO terms (Consortium., 2014); genes with multiple represented GO categories have been assigned to a representative one. The solid lines represent known direct interactions as annotated within the STRING (Franceschini et al., 2013) and BioGrid (Chatr-Aryamontri et al., 2013) databases with the dashed lines representing indirect interactions that are either between proteins found within same protein complex but not interacting directly or predicted interactions based on paralogs (Franceschini et al., 2013). The fully colored red nodes represent genes that have been directly implicated in ASD (from SFARI database (Banerjee-Basu and Packer, 2010) or Pinto et al. (Pinto et al., 2014)) whereas partially shaded nodes are genes associated with intellectual disability, FMRP (fragile X mental retardation) or within an ASD CNV but not, as yet, directly implicated in ASD (Banerjee-Basu and Packer, 2010; Pinto et al., 2014). Finally, GOrilla (http://cbl-gorilla.cs.technion.ac.il/) (Eden et al., 2009) was employed for the analysis of GO enrichment within each cluster of mouse microexons based on their dynamic PSI changes during neuronal differentiation (Table S4).

Analysis of microexon regulation during neuronal and myoblast differentiation

RNA-Seq data for six differentiation time points during neuronal differentiation from ES cells to cortical glutamatergic neurons were obtained from (Hubbard et al., 2013); only one time point DIV28 was used to represent fully mature neurons (see (Hubbard et al., 2013) for experimental details). ΔPSIs for each transition (T1 to T5) for neural microexons with sufficient read coverage across the six studied time points and a minimum $|\Delta PSI| \geq 25$ between ES cells (first time point) and fully mature glutamatergic neurons (last time point) were clustered using hclust function in R with default options (method = "complete"). ΔPSIs were visualized using enhanced heatmap in R (heatmap.2) from the gplots package. A manual tree cutoff was set and eight distinct clusters were defined based on their dynamic PSI values. For Figure S3A, cRPKMs for selected ES and

neural markers were represented by setting the maximum value for each gene across the six time points to 100 and the minimum to 0.

To analyze C2C12 myoblast differentiation (Figure S2D), we used mouse RNA-Seq data from (Trapnell et al., 2010), which includes four time points (-24 h, 60 h, 120 h and 168h of differentiation; Table S1). First, we obtained muscle-regulated microexons using the same definitions as per neural regulation, but comparing muscle and heart to non-muscle samples. Next, for muscle-regulated microexons with sufficient read coverage to confidently estimate PSIs in the four time points (n=41), we plotted ΔPSIs for each of the three transitions (T1 to T3) as a heatmap and performed an unsupervised clustering. In total, 22 events showed increased inclusion (ΔPSI>25) during myoblast differentiation, especially during T1. However, for 14/21 (66.7%) microexons with sufficient read coverage in glutamatergic neurons differentiated in vitro (DIV28), neurons display higher inclusion levels for the same microexons than detected in differentiated myotubes, and two showed similar inclusion levels (within ΔPSI<5).

Prediction of the impact of AS events on proteomes

Alternative sequences were first mapped to coding (CDS) or non-coding (UTRs) sequences based on Ensembl genome annotations. For those alternative sequences that had not been previously annotated, mapping was projected based on the information of the upstream exon. Then, sequences that contained in-frame stop codons (based on annotations and in-frame sequence translation from the upstream exon) or start codons (based on annotations) were flagged. AS events were predicted to generate alternative protein isoforms upon both inclusion and exclusion when they fall in the CDS, do not disrupt the reading frame (i.e. they have lengths multiple of three nucleotides), and do not contain in-frame stop codons predicted to trigger non-sense mediated decay (NMD; those stop codons that fall further than 50 nucleotides away from a downstream EEJ). Furthermore, if an alternative sequence is predicted to introduce an in-frame stop codon that would not trigger NMD but would generate a truncated protein >100 aminoacids shorter than the reference protein, it was also considered as an open reading frame (ORF)-disrupting event. For multiexonic cassette events (arrays of more than one alternative exon), NMD/ORF-disruption was assessed for all exons as a group based on their inclusion patterns in neural and non-neural samples (e.g. if one frame-shifting exon is downregulated in neural

samples and another upregulated, both the neural and non-neural isoforms can have intact reading frames). For ORF-disrupting events, two main categories were defined: ORF-disruption upon sequence inclusion or upon sequence exclusion. Therefore, depending on the neural inclusion pattern, we defined events that cause ORF disruption in neural samples, but not in other tissues ("ORF disruption in brain"), or the opposite, AS events that disrupt ORFs only in non-neural samples ("ORF disruption outside brain").

Comparison of protein features across exon types

To investigate protein features associated with each class of AS exon, we first mapped AS exons to annotated proteomes by translating each alternative exon and its neighboring exons in all 6 frames, and aligning them to the canonical UniProt sequence (Consortium., 2014). If the AS exon could not be mapped to UniProt proteins by this procedure, we repeated it for all protein-coding isoforms from Ensembl v63, starting with the longest isoform. When an AS exon is not identified within Ensembl or UniProt sequences, and to account for novel sequences identified by our approach, we combined the upstream (C1) and downstream (C2) exons and repeated the aforementioned procedure without the AS exon. If a match is identified, the AS exon is then translated in the correct frame (based on C1 information). Several features were investigated for mapped exons. Disorder was predicted using IUPred (Dosztányi et al., 2005) with a disorder cutoff of 0.4 (Davey et al., 2012). Putative linear motifs were predicted with ANCHOR (Mészáros et al., 2009); at least five residues in a row were required for a valid motif (value > 0.4), in line with the average length of linear motifs (Davey et al., 2012). Globular domains known to bind linear motifs and lipids were extracted from the ELM database (Dinkel et al., 2014) and Pfam (Finn et al., 2014), respectively. Known protein complexes were taken from the census of human soluble protein complexes (Havugimana et al., 2012), and degree for each protein within protein-protein interaction networks was obtained from BioGrid (Ellis et al., 2012). Absolute all-atom solvent-accessible surface areas (ASAs) for all monomeric structures contained in the Protein Data Bank in Europe (PDBe)(Gutmanas et al., 2014) and, when necessary, SWISS-MODEL database (Kiefer et al., 2009), were calculated using the PISA software, as contained in version 6.3.0 of the CCP4 structural analysis software package (Potterton et al., 2003); default parameters were used. Relative solvent-accessible surface areas (rASAs) were then calculated using a reference table of all-atom ASAs for the central amino acid

in Gly-X-Gly tripeptides (Miller et al., 1987). The mapping to UniProt was done using EBI SIFTS program (Velankar et al., 2013). Predicted solvent-accessible surface area scores were calculated using the NetSurfP version 1.1 (Petersen et al., 2009) and default assignments of accessibility were used.

In addition, we used another parallel approach to investigate alternative exon disorder and exon association with protein domains. First, we downloaded the annotated proteome from Ensembl v71 and mapped exonic genomic coordinates to protein coordinates exons using BioMart. Next, to include unannotated exons and microexons, we recreated novel protein isoforms by introducing the exonic sequence downstream of the upstream exon (C1), prioritizing the reference protein isoform (based on BioMart), if multiple C1-containing isoforms were available. In total, 11,584 human new protein isoforms were added to the Ensembl annotations using this approach. This expanded set of proteins was used to run *de novo* Disopred2 (to predict structural disorder, (Ward et al., 2004)), and Pfam (to identify protein domains, (Finn et al., 2014); only domains from the A module were used), both with default parameters.

For consistency, for all exon classes, only exons that would generate protein isoforms both when included and skipped (i.e. internal exons with lengths multiple of three nucleotides without in-frame stop codons) were analyzed.

Protein structure retrieval and modeling

In order to investigate the location of neural-regulated microexons in the protein structures (Figure S6A), we first searched for available structures in PDB (Gutmanas et al., 2014) containing these microexons. For this, we used BLASTX searches against the whole PDB using the concatenated sequence from C1, A and C2 exons for 301 neural-regulated microexons as queries. Hits with an E-value > 10 and with PDB chains not belonging to one of the mapped PDB identifiers were discarded. The remaining BLASTX alignments were checked to ensure no gaps were present in the region corresponding to the translated microexon sequence. In parallel, we mapped genomic exon coordinates onto Ensembl (v75) transcript isoforms using Biomart, and selected those isoforms that encode neural-regulated microexons. The corresponding microexon-containing Ensembl protein isoforms were then mapped to PDB identifiers using Biomart.

Reliable structures for 7 microexons were contained in PDB. Next, we modeled the structure of microexon-containing Ensembl proteins using the Phyre fold-recognition server (Kelley and Sternberg, 2009), with default parameters. For those structures that could be modeled, amino acid sequences were then recovered to check complete coverage of the microexon sequences plus four extra amino acids at each side, to ensure sequence specificity. After these filters, we obtained modeled structures for 47 additional neural-regulated microexons. Out of these 54 structures, 5 incomplete models were discarded upon manual inspection. PyMol version 1.5.0.4 (Schrödinger, LLC) was used to visualize protein structures using "select peptide, chain %", "select exon, peptide and resi $$-$$", "orient exon", and "zoom peptide" as default parameters, followed by manual adjustments for specific structures to maximize microexon visibility.

Evolutionary analysis of neural-regulated cassette exons

We analyzed three different aspects of evolutionary conservation of neural-regulated alternative exons (Irimia et al., 2009). First, to determine whether an alternative exon is conserved at the genomic level between human and mouse, we performed exon coordinate conversions using Galaxy LiftOver (from hg19 to mm9). Human exons with a LiftOver hit in mouse, and with associated AG (splicing acceptor) and/or GT (splicing donor) sites – to allow for non-exact lift-overs and a non-conserved alternative 5´ or 3´ splice site – in mouse were considered to be Genome-conserved. Second, if the orthologous exon is defined as alternative in both species (see "Alternative splicing definition and minimum read coverage"), AS of the exon was defined as conserved. Finally, if both orthologous exons have been defined as neural-regulated (see "Definition of neural-regulated alternative splicing"), they were considered to be conserved at the neural-regulatory level. This analysis was performed for all human AS exons (Figure 2D) and for the subset of neural-regulated human alone (Figure S2E). Comparisons of conservation between length classes (3-15 nt, 16-27 nt and >27 nt) were done using two-sided proportion tests.

To identify deeply conserved 3-15 nt microexons across vertebrates, we first selected those microexons that are neural-regulated in human and/or mouse. Next, we performed exon coordinate conversions using LiftOver as described above against chicken (galGal3), frog (xenTro3) and zebrafish (danRer7). In addition, we employed previously defined (Venkatesh et al., 2014) clusters of 1-to-1 orthologs for 10 vertebrate species using InParanoid, including

human, mouse, chicken, frog, zebrafish and elephant shark. For each ortholog in those species, we extracted microexons of the exact length as the mammalian ones. Orthology of each exon was then manually investigated based on C1, A and C2 sequences. Conserved microexons in each species along with associated data and sequences are shown in (Table S3).

To investigate the conservation of exonic sequences and their flanking intronic regions (150 nucleotides upstream and downstream of the exon), we used human phastCons (Siepel et al., 2005) data from alignments of 46 placental mammals (46way.placental), downloaded from UCSC (http://genome.ucsc.edu/). Only exons conserved at the genomic level between human and mouse were used for this analysis.

Analysis of splicing factor regulation of microexons

We followed similar procedures as those recently described in our global analysis of a conserved nSR100 regulated exon network in mammals (Raj et al., 2014).

- AS profiling of splicing factor-deficient or -overexpressing cells: we used available RNA-Seq data from splicing factor knock-downs (RBFOX1 in human neural precursor cells (Fogel et al., 2012), MBNL1 and MBNL2 in human HeLa cells (Han et al., 2013), ESRP1 in human PNT2 cells (Dittmar et al., 2012), and Srrm4/nSR100 and Ptbp1 in mouse N2A cells (Raj et al., 2014)), knock-outs (Ptbp2 in mouse cortex (P1 stage) and whole embryonic brain (E18.5 stage) (Li et al., 2013), and Rbfox1 in mouse (Lovci et al., 2013)), or overexpression (nSR100 in human 293T cells (Raj et al., 2014)), and their associated controls, to identify AS exons and microexons whose inclusion levels are affected by experimental manipulation of the expression of the splicing factors. We used two levels of variation: $15< |\Delta PSI| <25$ and $|\Delta PSI| > 25$.

- Analysis of PAR-iCLIP signals: We used PAR-iCLIP data obtained from flag epitope-tagged nSR100-expressing human 293T cells (Raj et al., 2014). Three biological replicates were analyzed. Reads were mapped to the human genome (hg19) using bowtie with parameters -m 5 -v 2 -a --best –strata, after barcode sequences and adaptors were trimmed from the 3´ and 5´ ends of the reads, respectively. In order to obtain nucleotide-level resolution of protein-RNA crosslinking, PAR-iCLIP reads mapped to the genome were reduced to their first coordinate, and

the number of tags overlapping each genomic position was calculated. To account for differences in library size and moderate biases in PAR-iCLIP signals from differential expression of binding substrates, the resulting profiles were normalized for each interval *i* comprising an exon and the region 200 nt upstream and downstream by multiplying with $(l_i + \mu_l) / (N \cdot (r_i + \mu_r))$, where l_i is the interval length, μ_l the mean interval length, N the number of reads in the library in million, r_i the number of reads mapping to interval i, and μ_r the mean number of reads per interval. For alignment plots (Figure 4B), the three replicates were averaged, and only AS events with enough read coverage (see above) in the RNA-Seq sample from nSR100-expressing 293T cells were used. These included neural microexons (3-27 nt), long neural exons (>27 nt) and a subset of non-neural AS exons matched for PSI against all neural exons. Profiles for all intervals were aligned with respect to the 3' or 5' splice site (for left and right half of the panel in Figure 4B, respectively), alignments across all intervals in each group were averaged while accounting for the number of aligned regions at each position, and smoothened using an 11-nt running mean centered on each position.

- Cumulative Distribution plots analyzing nSR100 binding sites: For each upstream intron, the sequence from 3´ to 5´ starting from the AG dinucleotide (splice site acceptor) was scanned, and the position of the first occurrence of a TGC (a previously-defined nSR100 binding motif (Nakano et al., 2012; Raj et al., 2014)) was scored. These values were then plotted cumulatively for each set of exons of interest (Figure 4C).

- Splice Site Strength Calculations: Splice site scores were calculated using score5.pl and score3.pl from (Yeo and Burge, 2004). This method uses a position weight matrix and calculates deviation from the consensus. For 5´ splice sites, three exonic and six intronic positions surrounding the exon-intron junction were analyzed, and for the 3´ splice site, twenty intronic and three exonic positions were analyzed.

Comparison of autistic and control brains

Autism tissue was acquired from two brain banks - the Eunice Kennedy Shriver NICHD Brain and Tissue Bank for Developmental Disorders, the Autism Tissue Program, and the Harvard Brain Tissue Resource Center. In total, we analyzed 24 autistic individuals and 24 controls

matched by age and gender, and factors related to RNA quality and technical variables (RIN, batch effects, and sequencing depth) were used as covariates. Brain samples were dissected retaining grey matter from all cortical layers, isolating 50-100mg of tissue across the superior temporal gyrus (Brodmann areas ba41/42/22). RNA was isolated using the miRNeasy kit (Qiagen). Ribosomal RNA was depleted from 2ug total RNA with the Ribo-Zero Gold kit (Epicentre). Remaining RNA was then size selected with AMPure XP beads (Beckman Coulter) and resuspended in 8.5 ul of Illumina resuspension buffer and 8.5 ul of 2x EPF buffer. Subsequent steps followed the Illumina TruSeq protocol (starting at page 84 of the sample prep v2 guide, no changes). After this protocol was followed, libraries were quantified with the Quant-iT PicoGreen assay (Life Technologies) and validated on an Agilent 2200 TapeStation system. Libraries were pooled to multiplex 24 samples per lane using Illumina barcodes, and each pool was sequenced six times on a HiSeq2000/2500 instrument using high output mode with standard chemistry and protocols to achieve an average depth of 64M reads for 50bp paired end reads (~32M fragments on average).

For gene-level differential expression analysis, reads were aligned to the human genome (hg19) with TopHat2 (Kim et al., 2013) and quantified with HT-seq Counts using a union-exon model from Gencode v18 (Anders et al., 2014). Only genes with Fragments Per Kilobase Of Exon Per Million Fragments Mapped (FPKM) > 1 in 80% of samples were included in further analyses. Differential expression was assessed using a linear model with log2(FPKM) as the outcome, and diagnosis, age, sex, RIN, brain bank, a surrogate variable for sequencing depth, and a surrogate variable for sequencing 3' bias and GC content as covariates. A set of 12 samples were selected based on the ASD samples showing the greatest global differential expression signature among all ASD samples, and the control samples showing a signal that was closest to the median of all controls. Importantly, no knowledge of splicing changes went into this sample selection.

To investigate misregulation of nSR100 gene expression between these selected ASD and control samples, we performed differential expression analysis using the linear model as described above on the subset of 12 vs. 12 samples. Beta values (log2 transformed fold-changes associated with predictors) and p values for ASD status are reported, and multiple comparisons were accounted for using the positive FDR (q value). For comparison, requiring equal or lower p and beta values

than those of nSR100, 217 genes showed similar or stronger down-regulation at the gene expression level in ASD versus control samples.

For splicing analysis, the selected samples were grouped into three sets of four individuals each based on age similarity (Table S1). For an exon to be considered as misregulated in ASD, we required an |ΔPSI| of at least 10 between the average PSI of ASD and control groups. To study whether nSR100 regulation was associated with AS misregulation in ASD (Figure 7C), we divided exons between nSR100-regulated (ΔPSI>25 between nSR100-overexpressing and control 293T cells) and non-regulated (|ΔPSI|<5) and calculated the fraction within each group that showed an average |ΔPSI| > 10 in ASD vs. control groups. Furthermore, for those microexons with sufficient read coverage to derive confident PSIs in at least 9 ASD and 9 control individual brain samples, we calculated the correlation coefficients between their PSIs and nSR100 expression across samples (Figure 7D). For this analysis, microexons were divided by divided between nSR100-regulated (ΔPSI>25 between nSR100-overexpressing and control 293T cells) and non-regulated (|ΔPSI|<10).

RT-PCRs for a subset of misregulated microexons (Figure S7B) and qRT-PCRs to evaluate nSR100 expression (Figure S7D) were performed using total RNA from 4 representative ASD individual and 3 controls ("Exp. Validation" column in Table S1, and asterisks in Figure S7C). nSR100 values from qRT-PCRs were normalized against the housekeeping gene GAPDH; in addition, two other housekeeping genes (EEF2 and RPS25) with high expression levels and minimal differences between ASD and control individuals were used for independent normalizations.

Structural analysis of the APBB1 microexon

We focused on a deeply conserved (Table S3) 6-nt microexon located in a loop within the phosphotyrosine binding 1 (PTB1) domain of APBB1. Since there is no available crystal structure for APBB1-PTB1 in complex with protein ligands, and the molecular details of the interacting interface between APBB1-PTB1 and KAT5 are still controversial (Radzimanowski et al., 2008a), we undertook an alternative approach that relies on structural alignments. Previous studies have shown that different PTB domain subfamilies share a structurally similar interacting

interface (Uhlik et al., 2005). We thus used the Dali server (Holm and Rosenström, 2010) to identify structures that have been resolved in the presence of a protein ligand and that most closely resembles the APBB1-PTB1 domain. Based on this information, we selected the APBB1-PTB2 bound to a 35 amino acid amyloid precursor protein (APP) fragment (root-mean-square deviation (RMSD) of 2.4Å) to perform the structural analyses (better Dali hits, up to 1.9Å, consisted only of small peptic ligands, smaller than 15 amino acids). The structures of the unbound form of the APBB1-PTB1 domain (PDB ID: 3D8F) (Radzimanowski et al., 2008a) and the APBB1-PTB2 in complex with APP (PDB ID: 3DXC) (Radzimanowski et al., 2008b) were then structurally aligned using the Pymol Molecular Graphics System (Schrödinger, LLC). Next, to determine the presence of the microexon in relation to the predicted APBB1-PTB1 binding interface, we utilized the CSU algorithm (Sobolev et al., 1999) to analyze the atomic distances between the superimposed APBB1-PTB1 domain and the APP protein fragment. Rigid structural alignment between the unbound APBB1-PTB1 domain and the APBB1-PTB2 in complex with APP protein fragment showed that the Glu residue encoded by the microexon is located ~5 Å from the APP protein fragment. This thus suggests that microexon skipping may reduce binding of interaction partners, including KAT5, by affecting directly the APBB1-PTB1 binding interface.

LUMIER assay

HEK-293T cells were transiently transfected using Polyfect (Qiagen) with *Renilla* Luciferase (RL)-tagged Apbb1, with or without inclusion of the alternative microexon, together with 3Flag-tagged Kat5 or App. Two days after transfection cells were lysed in 0.5% TNTE (50mM Tris at pH7.4, 150mM sodium chloride, 1mM EDTA, 0.5% Triton X-100), in the presence of a cocktail of protease and phosphatase inhibitors. Lysates were transferred to a 96-well plate coated with anti-Flag M2 antibody (Sigma). Plates were incubated at 4°C for an hour, after which plates were washed 5 times with 0.1% TNTE. Luminescence in each well was measured with an Envision (Perkin Elmer) plate reader using *Renilla-Glo*® Luciferase Assay System (Promega). Normalized LUMIER intensity ratio (NLIR) values were calculated as described previously (Ellis et al., 2012).

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq – A Python framework to work with high-throughput sequencing data. bioRxiv.

Banerjee-Basu, S., and Packer, A. (2010). SFARI Gene: an evolving database for the autism research community. Dis Model Mech *3*, 133-135.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.* (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587-1593.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Frey, B.J., Irimia, M., and Blencowe, B.J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res, pii: gr.177790.177114.

Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., *et al.* (2013). The BioGRID interaction database: 2013 update. Nucleic Acids Res *41*, D816-823.

Consortium., U. (2014). Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res *42*, D191-198.

Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., *et al.* (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature *492*, 59-65.

Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012). Attributes of short linear motifs. Mol Biosyst *8*, 268-281.

Dinkel, H., Van Roey, K., Michael, S., Davey, N.E., Weatheritt, R.J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kuban, M., *et al.* (2014). The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res *42*, D259-266.

Dittmar, K.A., Jiang, P., Park, J.W., Amirikian, K., Wan, J., Shen, S., Xing, Y., and Carstens, R.P. (2012). Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. Mol Cell Biol *32*, 1468-1482.

Dosztányi, Z., Csizmók, V., P., T., and Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol *347*, 827-839.

Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. PLoS Comput Biol *5*, e1000376.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists. BMC Bioinformatics *10*, 48.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., *et al.* (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell *46*, 884-892.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* (2014). Pfam: the protein families database. Nucleic Acids Res *42*, D222-230.

Fogel, B.L., Wexler, E., Wahnich, A., Friedrich, T., Vijayendran, C., Gao, F., Parikshak, N., Konopka, G., and Geschwind, D.H. (2012). RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. Hum Mol Genet *21*, 4171-4186.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., *et al.* (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res *41*, D808-815.

Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Gore, S.P., *et al.* (2014). PDBe: Protein Data Bank in Europe. Nucleic Acids Res *42*, D285-291.

Han, H., Irimia, M., Ross, P.J., Sung, H.K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., *et al.* (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. Nature *498*, 241-245.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., *et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068-1081.

Holm, L., and Rosenström, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545-549.

Hornbeck, P.V., Chabra, I., Kornhauser, J.M., Skrzypek, E., and Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics *4*, 1551-1561.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Hubbard, K.S., Gut, I.M., Lyman, M.E., and McNutt, P.M. (2013). Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. F1000Res *2*, 35.

Irimia, M., and Blencowe, B.J. (2014). An atlas of alternative splicing profiles. In preparation.

Irimia, M., Rukov, J.L., Roy, S.W., Vinther, J., and Garcia-Fernandez, J. (2009). Quantitative regulation of alternative splicing in evolution and development. Bioessays *31*, 40-50.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc *4*, 363-371.

Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M., *et al.* (2012). 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. Nat Struct Mol Biol *19*, 1037-1043.

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Res *37*, D387-392.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.

Labbe, R.M., Irimia, M., Currie, K.W., Lin, A., Zhu, S.J., Brown, D.D., Ross, E.J., Voisin, V., Bader, G.D., Blencowe, B.J., *et al.* (2012). A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. Stem Cells *30*, 1734-1745.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Li, Q., Zheng, S., Han, A., Lin, C.H., Stoilov, P., Fu, X.D., and Black, D.L. (2013). The splicing regulator PTBP2 controls a program of embryonic splicing required for neuronal maturation. Elife *3*, e01201.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., *et al.* (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol *20*, 1434-1442.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162-1164.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One *5*, e13984.

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS Comput Biol *5*, e1000376.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. (1987). Interior and surface of monomeric proteins. J Mol Biol *196*, 641-656.

Nakano, Y., Jahan, I., Bonde, G., Sun, X., Hildebrand, M.S., Engelhardt, J.F., Smith, R.J., Cornell, R.A., Fritzsch, B., and Bánfi, B. (2012). A mutation in the Srrm4 gene causes alternative splicing defects and deafness in the Bronx waltzer mouse. PLoS Genet *8*, e1002966.

Parikshak, N., and Geschwind, D.H. (2014). In preparation.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol *9*, 51.

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods *8*, 785-786.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., *et al.* (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet *94*, 677-694.

Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003). A graphical user interface to the CCP4 program suite. Acta Crystallogr D Biol Crystallogr *59*, 1131-1137.

Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., and Ouzounis, C.A. (2000). CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinformatics *16*, 915-922.

Radzimanowski, J., Ravaud, S., Schlesinger, S., Koch, J., Beyreuther, K., Sinning, I., and Wild, K. (2008a). Crystal structure of the human Fe65-PTB1 domain. J Biol Chem *283*, 23113-23120.

Radzimanowski, J., Simon, B., Sattler, M., Beyreuther, K., Sinning, I., and Wild, K. (2008b). Structure of the intracellular domain of the amyloid precursor protein in complex with Fe65-PTB2. EMBO Rep *9*, 1134-1140.

Raj, B., Irimia, M., Braunschweig, U., Sterne-Weiler, T., O'Hanlon, D., Yuan-Lin, Z., Chen, I.G., Easton, L., Ule, J., Gingras, A.C., *et al.* (2014). Global regulatory mechanism underlying the activation of an exon network required for neurogenesis. Mol Cell *In press*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol *3*, Article3.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. Bioinformatics *15*, 327-332.

Sofueva, S., Yaffe, E., Chan, W.C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J *32*, 3119-3129.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van, B.a., M.J, Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Uhlik, M.T., Temple, B., Bencharit, S., Kimple, A.J., Siderovski, D.P., and Johnson, G.L. (2005). Structural and evolutionary division of phosphotyrosine binding (PTB) domains. J Mol Biol *345*, 1-20.

Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res *41*, D483-489.

Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B., Ohta, Y., Flajnik, M.F., Sutoh, Y., Kasahara, M., *et al.* (2014). The elephant shark genome provides unique insights into gnathostome evolution. Nature *505*, 174-179.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol *337*, 635-645.

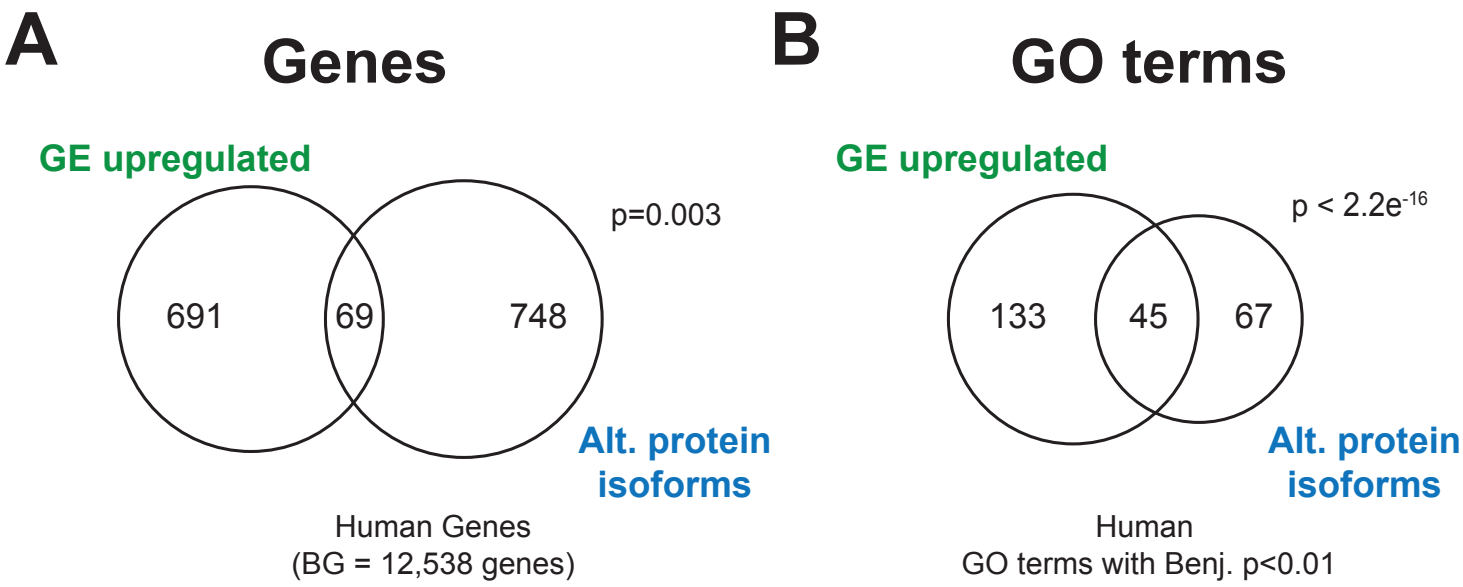
Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., *et al.* (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res *38*, W214-220.

Wu, J., Anczuków, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res *41*, 5149-5163.

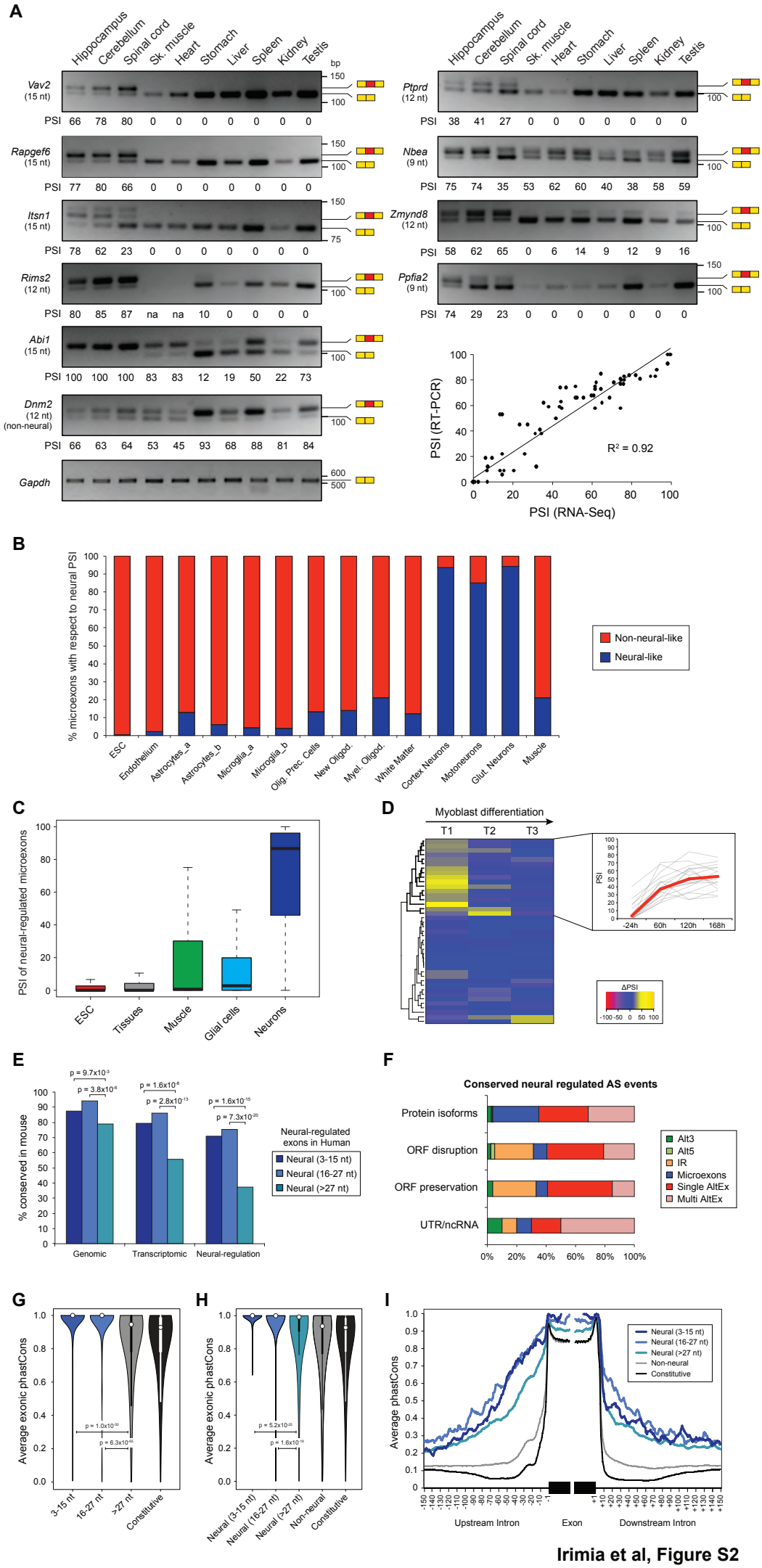
Yeo, G.W., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., *et al.* (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J Neurosci *34*, 11929-11947.

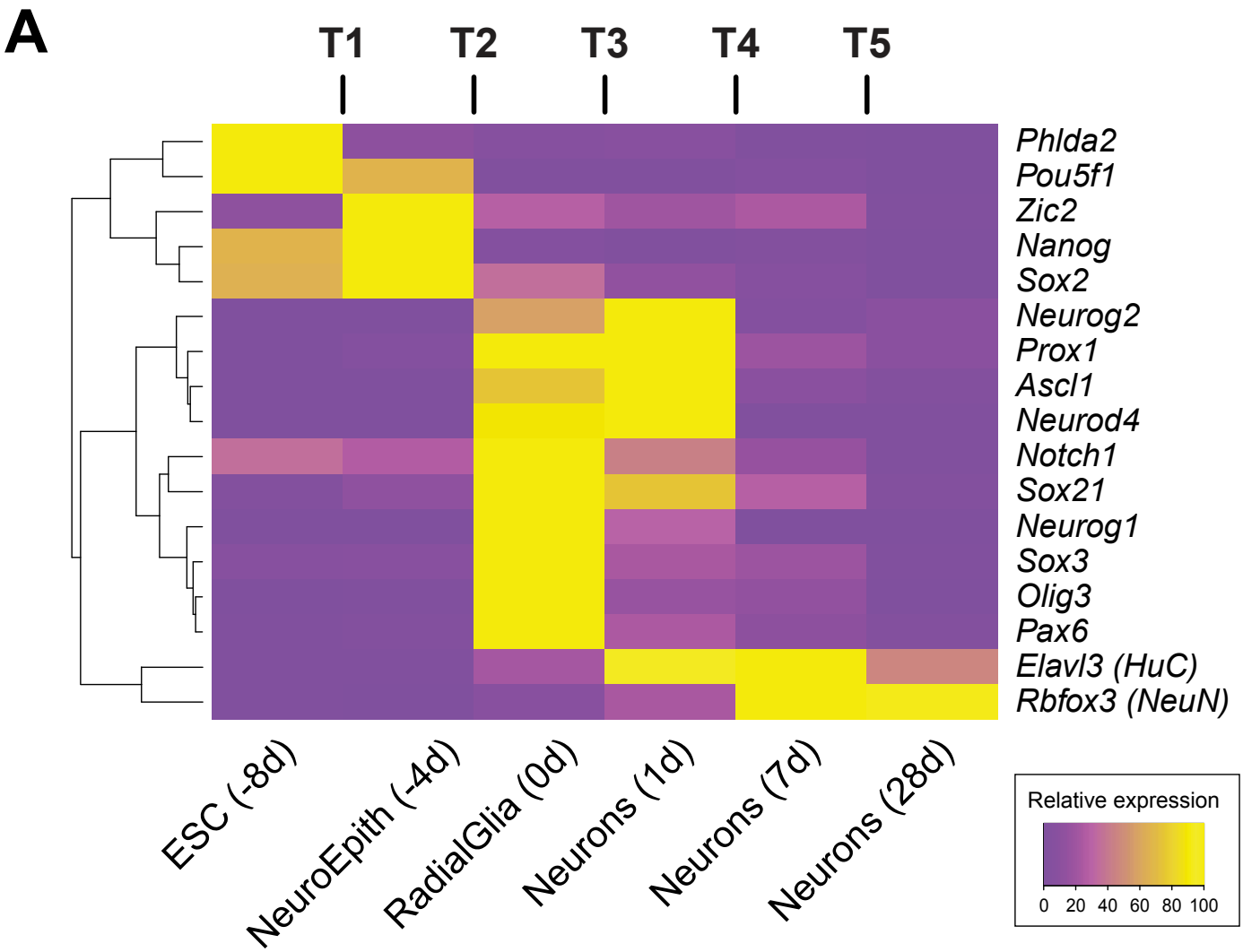
Zhang, Z., Pinto, A.M., Wan, L., Wang, W., Berg, M.G., Oliva, I., Singh, L.N., Dengler, C., Wei, Z., and Dreyfuss, G. (2013). Dysregulation of synaptogenesis genes antecedes motor neuron pathology in spinal muscular atrophy. Proc Natl Acad Sci USA *110*, 19348-19353.

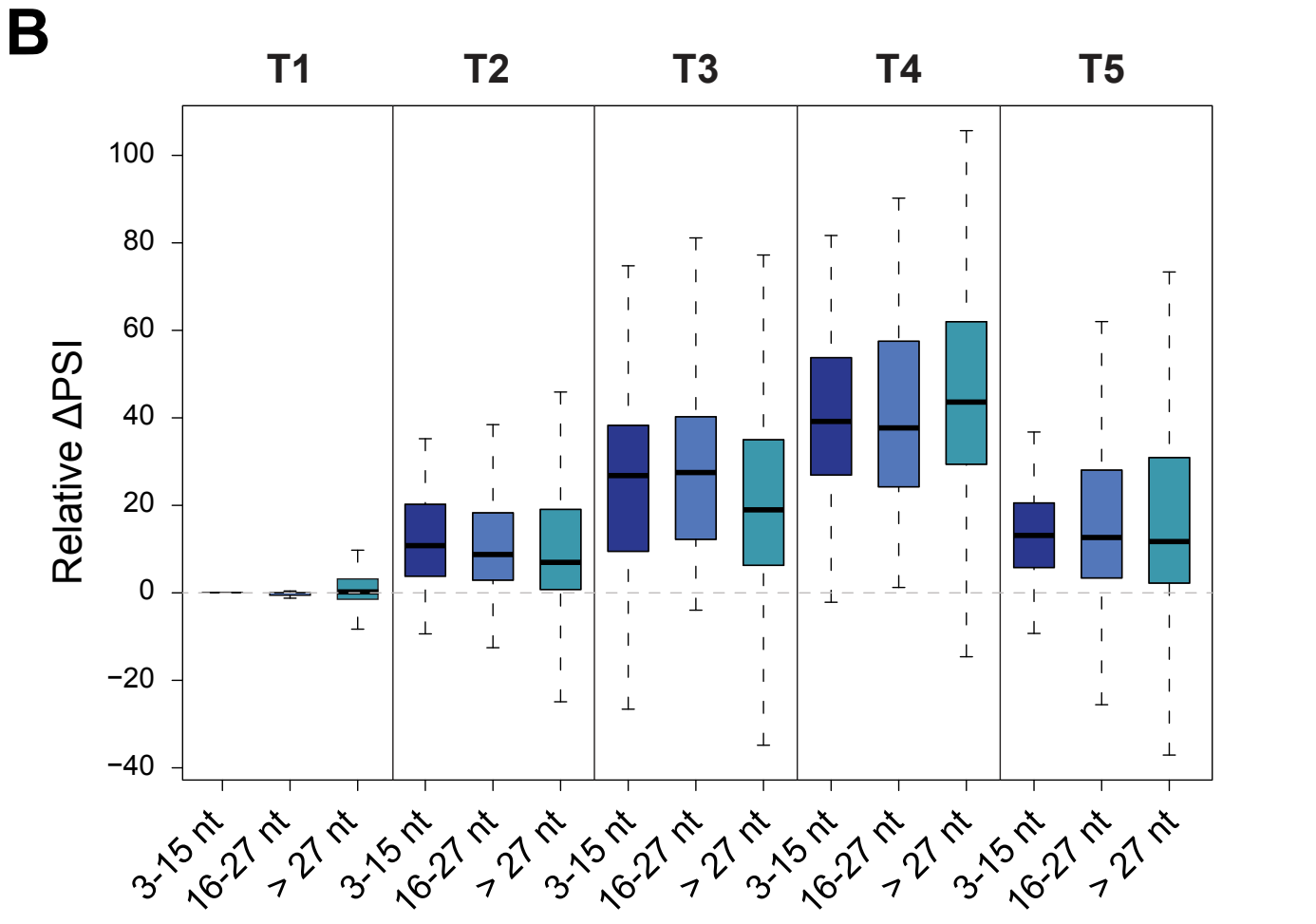


Irimia et al, Figure S1

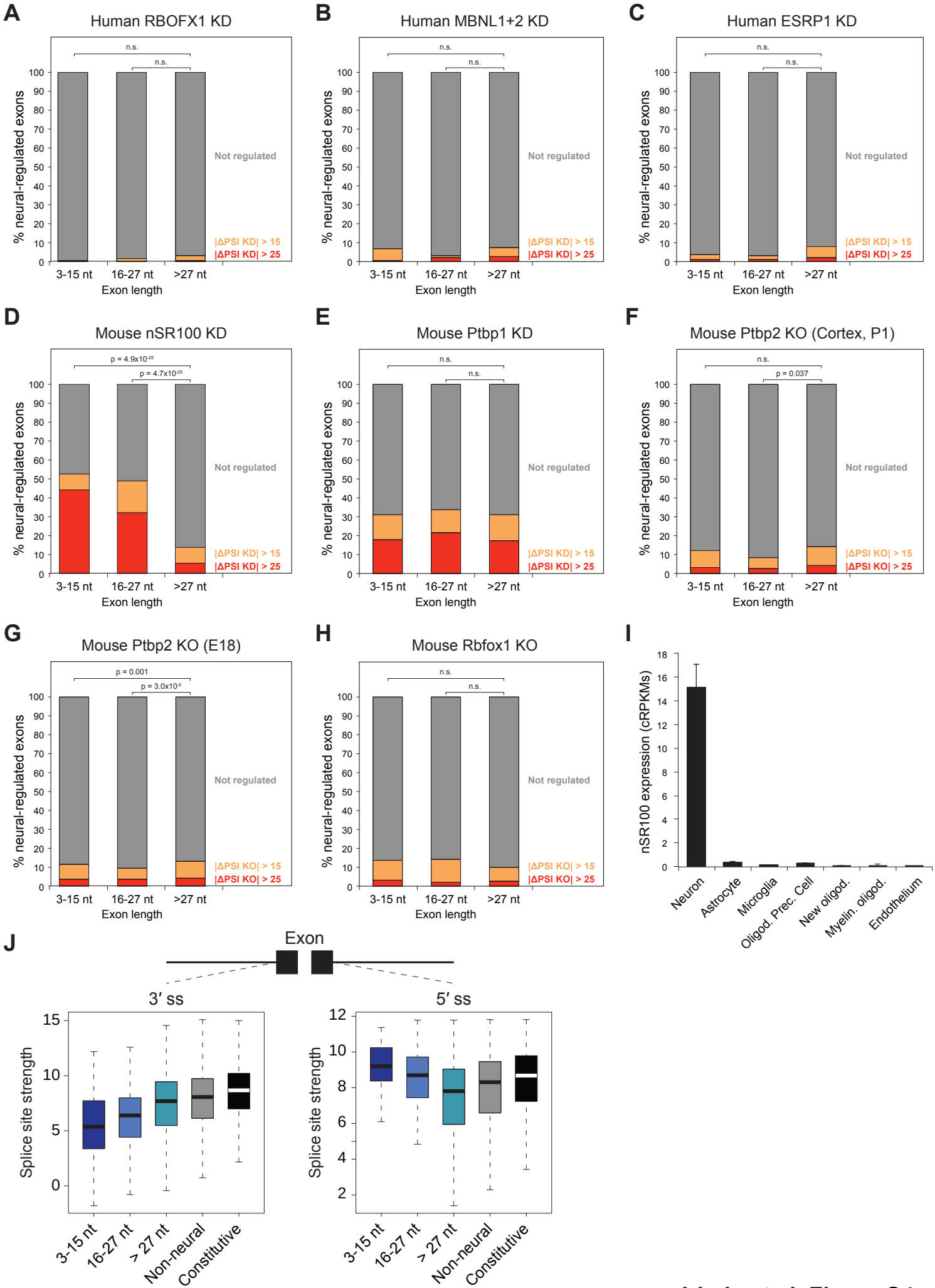


Irimia et al, Figure S2

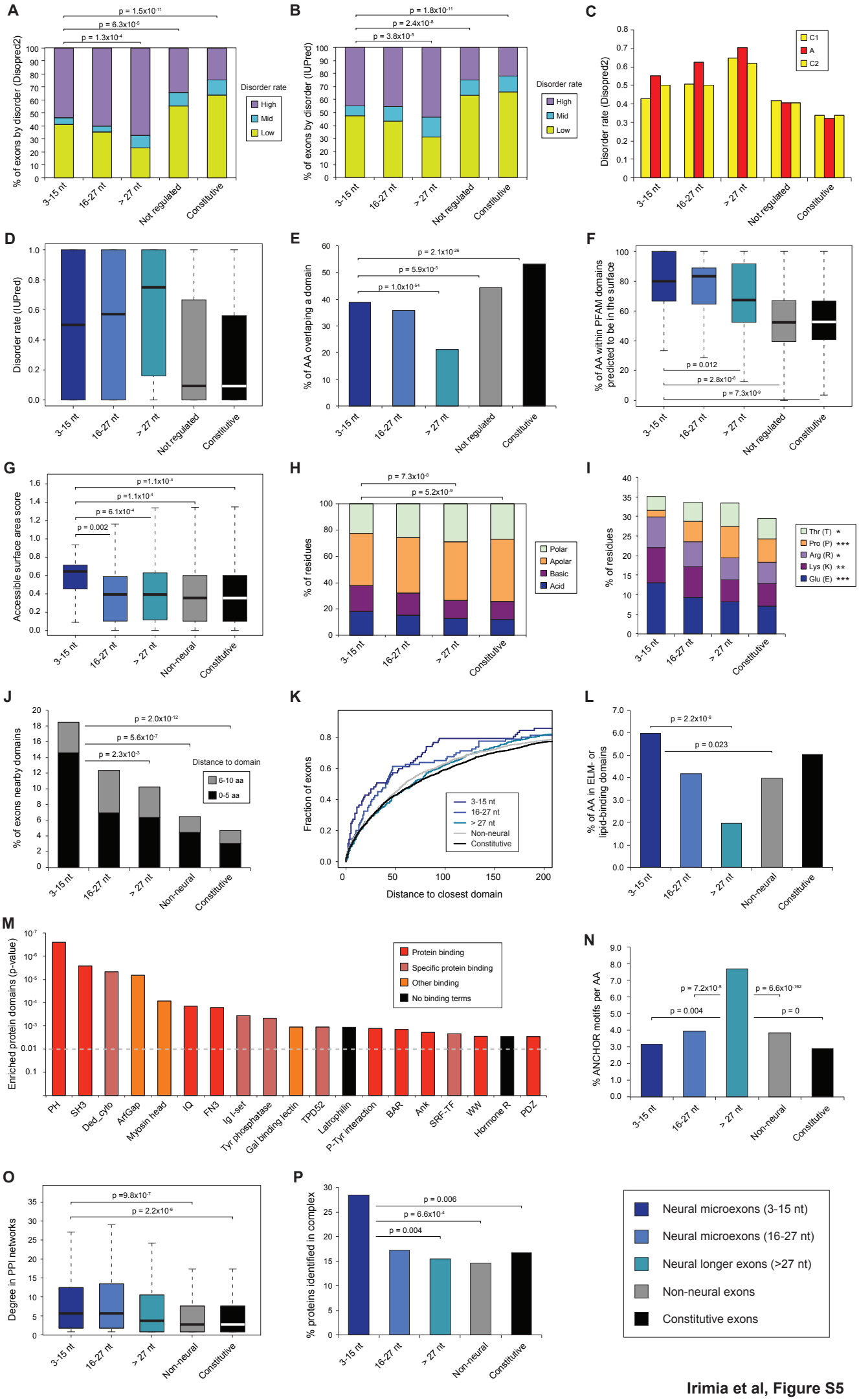




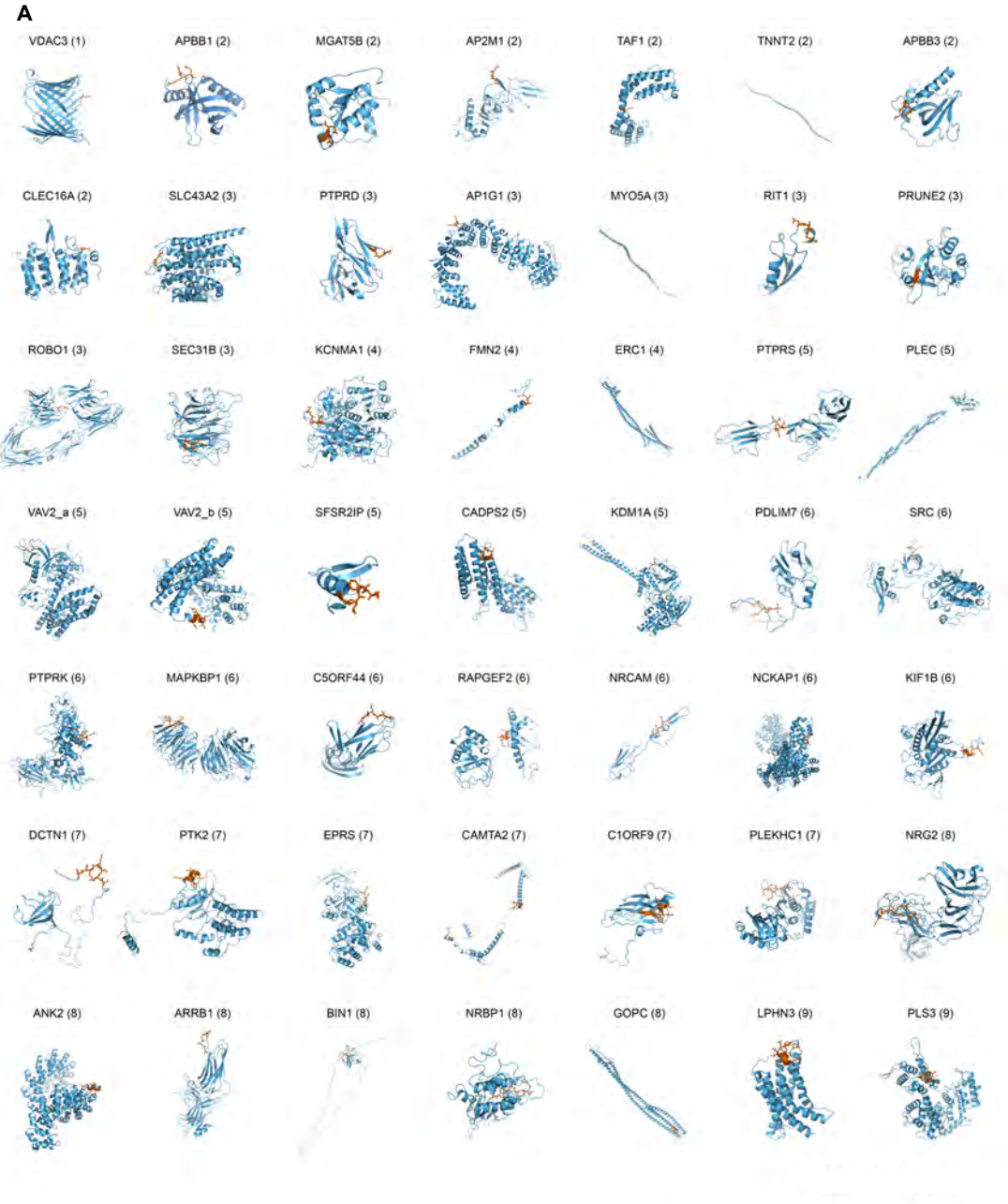
Irimia et al, Figure S3

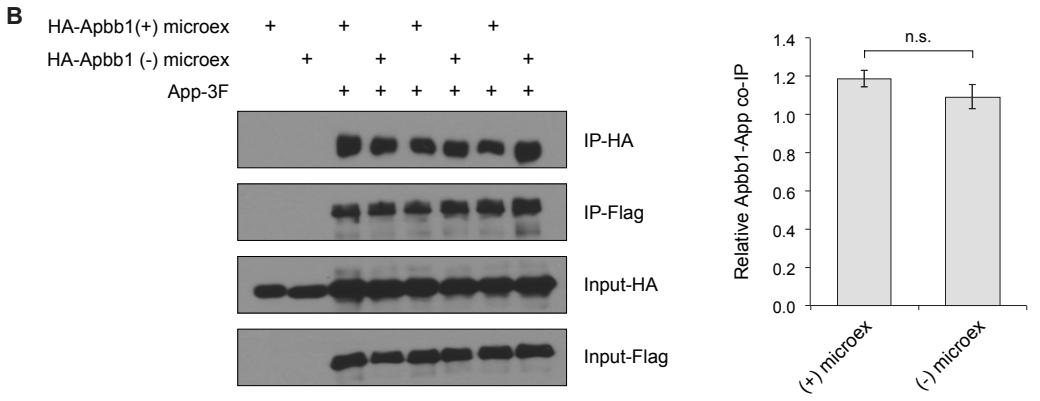


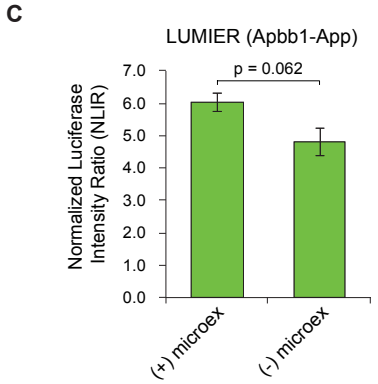
Irimia et al, Figure S4

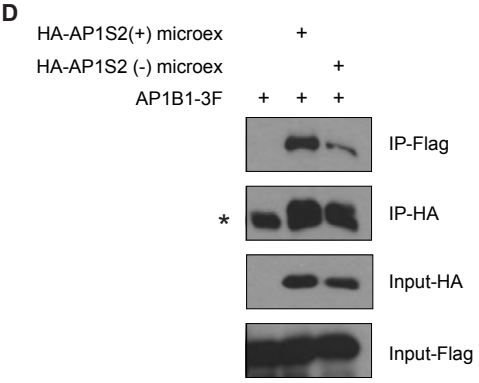


Irimia et al, Figure S5

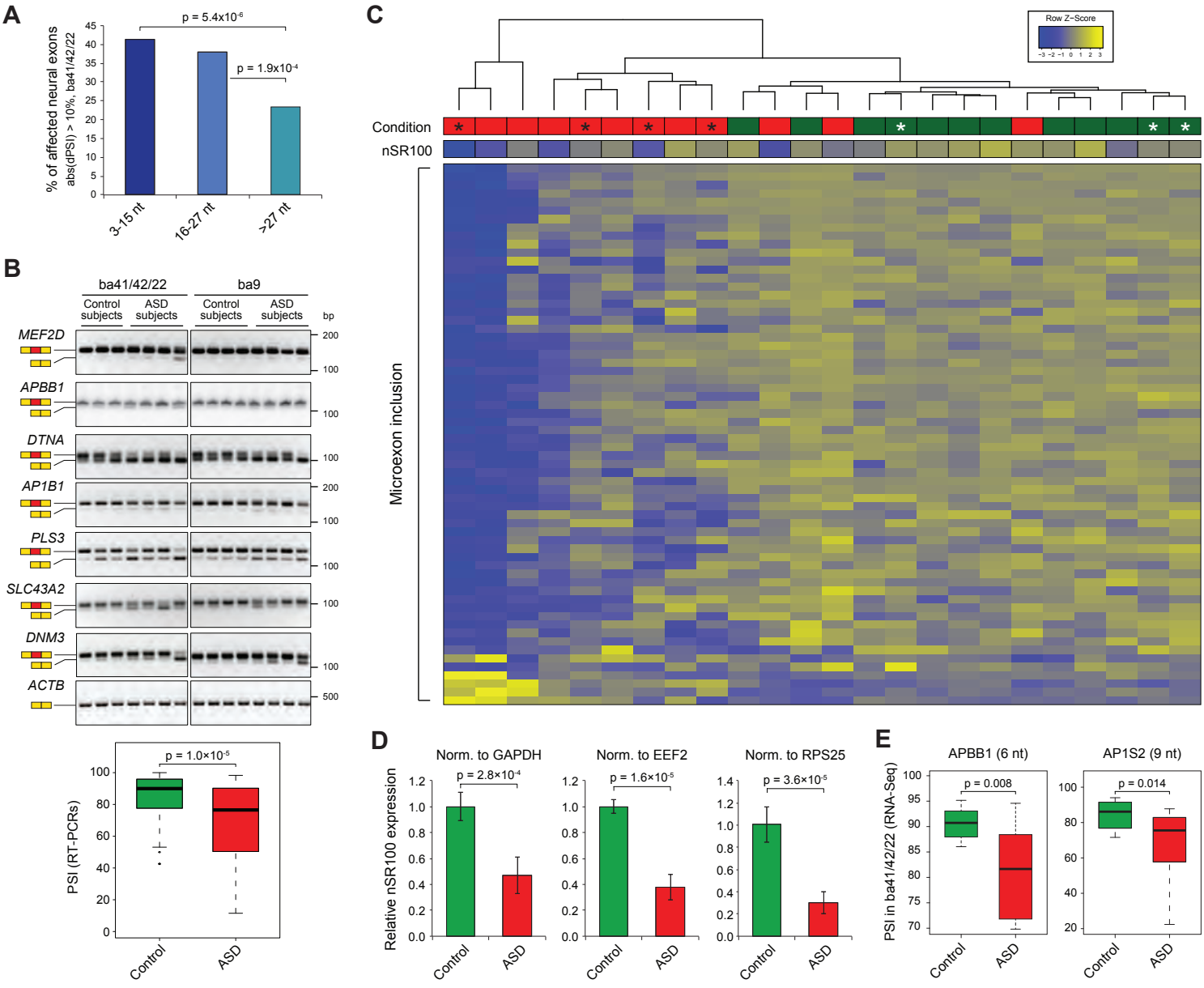








Irimia et al, Figure S6



Irimia et al, Figure S7