3 Chapter III: Microexon quantitative analyses across mouse brain development and visual cortex

Collaboration note

All of the work shown in this Chapter will be appended to some results from Chapter II to publish a manuscript under preparation, in which I will be the leading author and Eric Miska and Martin Hemberg will be the corresponding authors. Moreover, the results and figures here presented partially correspond to the current last version of the results that we are planning to submit within a month from this thesis submission date. I produced all the code necessary to carry the complete data analysis and data visualization here presented, but the overall product was only possible to collaborative efforts carried out by Roberto Munita, Ilias Georgakopoulos-Soares, Hugo Fernandez[4], Emmanouil Metzakopian[5], Maria Estela Andres[6].

3.1 Introduction

Even though the first reports of the highly neuron-specific microexon inclusion events date several decades ago (Santoni et al., 1989; Wiestler and Walter, 1988), only recent genome-wide analyses of microexons have enabled to uncover the landscape of neuronal microexon splicing events (Irimia et al., 2014; Li et al., 2015). These analyses have shown that microexons are a highly conserved and regulated network of neuronal events, which modify a wide range of neuronal proteins involved in neurogenesis and axonogenesis, synapse, kinase activity, vesicle transport and cytoskeleton regulation.

Quantitative analyses of microexon inclusion have shown clusters of coordinated microexon inclusion events that are progressively included through *in vitro* neuronal

[4] Postdoctoral fellow at UK Dementia Research Institute, Department of Clinical Neurosciences, University of Cambridge. Cambridge, UK

[5] Group leader at UK Dementia Research Institute, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

[6] Group leader at the Department of Cellular and Molecular Biology, Faculty of Biological Sciences, Pontificia Universidad Católica de Chile, Santiago, Chile.

differentiation (Irimia et al., 2014). These microexon splicing events are largely regulated by the combinatorial effects of a range of RBPs, such as SRRM4, RPBOX1 and PTB1 (Gonatopoulos-Pournatzis et al., 2018; Irimia et al., 2014; Li et al., 2015). Experimental knockdown and overexpression of SRRM4 have been shown to have large effects over neuronal cassette exon inclusion and have functional consequences for neurite outgrowth and interfere with neuronal differentiation process (Calarco et al., 2009; Raj et al., 2011, 2014). The generation of knockout SRRM4 mice showed that the loss of this protein factor results from impairments of the central and peripheral nervous systems, affecting neurite outgrowth, cortical layering and axon guidance (Quesnel-Vallières et al., 2015). Even though SRRM4 affects a wide range of alternative splicing events, microexons are the main group that is affected by SRRM4 absence and the re-establishment of *wild-type* PSI levels of a single microexon at *UNC13B* gene was shown to be sufficient to rescue a neuritogenesis defects induced by SRRM4 absence (Quesnel-Vallières et al., 2015). Together these results demonstrate the key role of microexon inclusion for normal neuritogenesis .

Since clusters of microexons have been shown to be progressively included during *in vitro* neuronal differentiation (Irimia et al., 2014), I hypothesized that there are groups of microexons that are differentially included during mouse embryonic development and that they have a wide range of effects on neuronal protein functions. The massive amount of data that is currently available in public repositories enabled unprecedented access to transcriptome complexity, however microexons cannot be efficiently detected when standard tools to process RNA-seq data are used. Therefore the processing of raw available RNA-seq experiments using methods that can reliably identify and quantify microexons are necessary to explore their tissue-specific patterns and dynamic splicing changes during developmental time. Moreover, since neuronal tissues such as the brain cortex are particularly diverse in terms of cell-types, the integration of scRNA-seq data has the potential to provide a detailed map of microexon splicing changes across brain cell-types. Thus, in this chapter I used MicroExonator to process a large set of bulk and single cell RNA-seq experiments in order to explore microexon inclusion patterns during mouse embryonic development and across cortical neuronal subtypes.

3.2 Results

3.2.1 Microexon inclusion changes dramatically over mouse embryonic development

To investigate how microexon inclusion patterns change during mouse development, I analysed 271 RNA-seq datasets generated by the ENCODE consortium (ENCODE Project Consortium, 2004). These RNA-seq data originate from 17 different tissues, (including forebrain, hindbrain, midbrain, neural tube, adrenal gland, heart, and skeletal muscle) across 7 different embryonic stages (ranging from E10.5 to E16.5), early postnatal (P0) and early adulthood (8 weeks). In addition, I analysed 18 RNA-seq experiments from mouse cortex across nine different time points; embryonic development (E.14.5 and E16.5), early postnatal (P4, P7, P17, P30), and older (4 months and 21 months) (Weyn-Vanhentenryck et al., 2018). Using the annotations provided by GENCODE and VastDB I detected 2,966 microexons in total, and I quantified their inclusion by calculating PSI values for each mouse sample. As some microexons were detected in lowly expressed genes, I only retained microexons whose inclusion or exclusion was supported by >4 reads in >10% of the samples, and this resulted in 2,557 microexons.

To characterize the splicing patterns I performed dimensionality reduction using probabilistic principal component analysis (PPCA) (Roweis, 1998; Tipping and Bishop, 1999), and I identified three components that together explain 79.4% of the total PSI variance across samples (Fig 3.1a-b). The first principal component (PC1) accounts for 56.9% of PSI variance and strongly correlates with embryonic developmental stage of neuronal samples measured as days post conception (DPC) between E10.5 and E14.5, suggesting a strong coordination of microexon splicing during brain embryonic development (Fig 3.1c). PC2 explains 16.7% of PSI variability and is exclusively related with muscular-specific microexon inclusion patterns that were detected in heart and skeletal muscle, suggesting muscle-specific microexon splicing patterns (Fig 3.1a). Finally, PC3 explains 6.2% of PSI variability and it is related to microexon alternative splicing changes in whole cortex postnatal samples, suggesting that microexon neuronal splicing keeps changing after birth, but to a lesser extent than during embryonic development (Fig 3.1b).

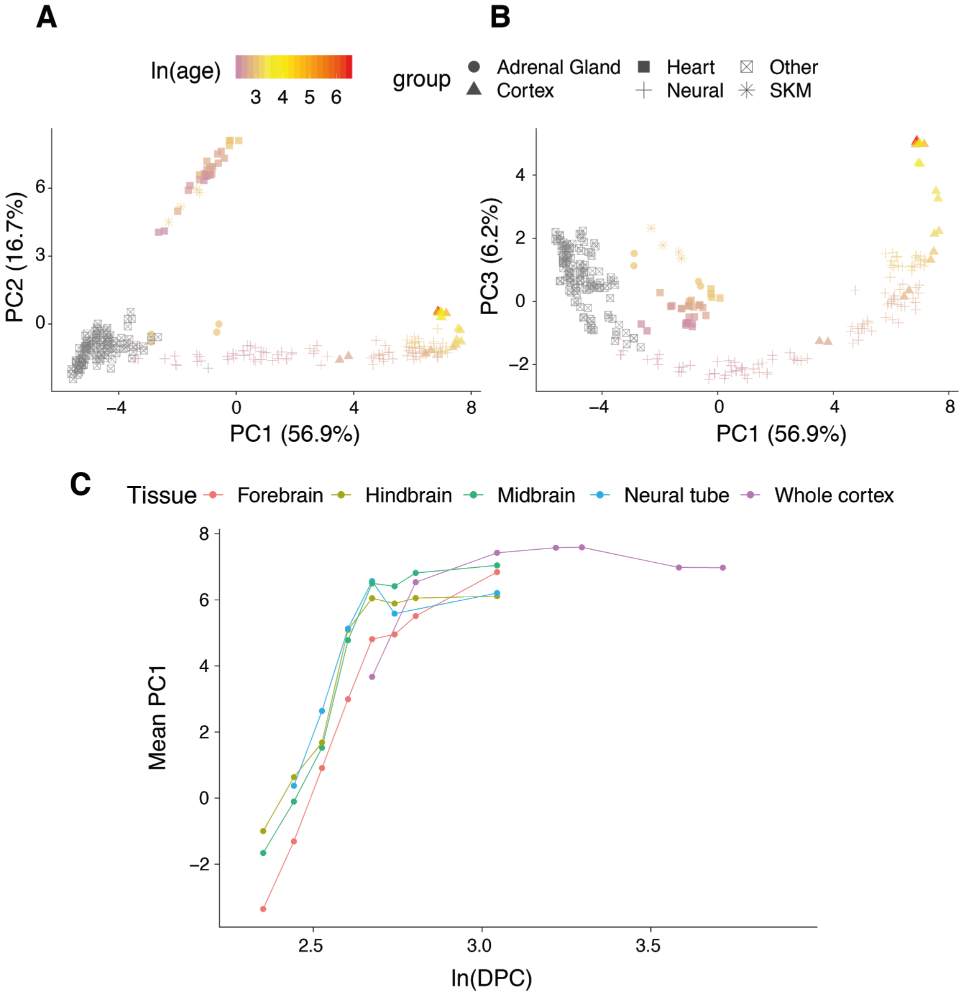

Figure 3.1. Microexon inclusion through mouse embryonic development. A. Dimensionality reduction using probabilistic principal component analysis of microexon PSI values across mouse embryonic and postnatal samples reveals correlation with developmental time for PC1. PC2 separates heart and SKM from other tissues. **B.** PC3 is correlated with developmental time of the postnatal brain samples. **C.** PC1 correspondence with embryonic developmental time, here expressed as log days post conception.

To further investigate tissue-specific microexon changes throughout development I performed biclustering of microexon PSI values from the different embryonic samples, and I obtained 24 microexon and 17 sample clusters (Fig 3.2a). Each of the sample clusters represents a combination of well defined subsets of tissues and embryonic states (Fig 3.2b). For example, samples corresponding to brain, heart, skeletal muscles (SKM) and adrenal gland (AG) form separate groups, with the only exception being E10.5 brain samples which clustered together with embryonic facial prominence and limb from E10.5 to E12.0. Consistent with the dimensionality reduction analysis, samples from the brain cluster preferentially by developmental time rather than by neuronal tissue, suggesting that microexon alternative splicing changes are greater between developmental stages than between brain regions.

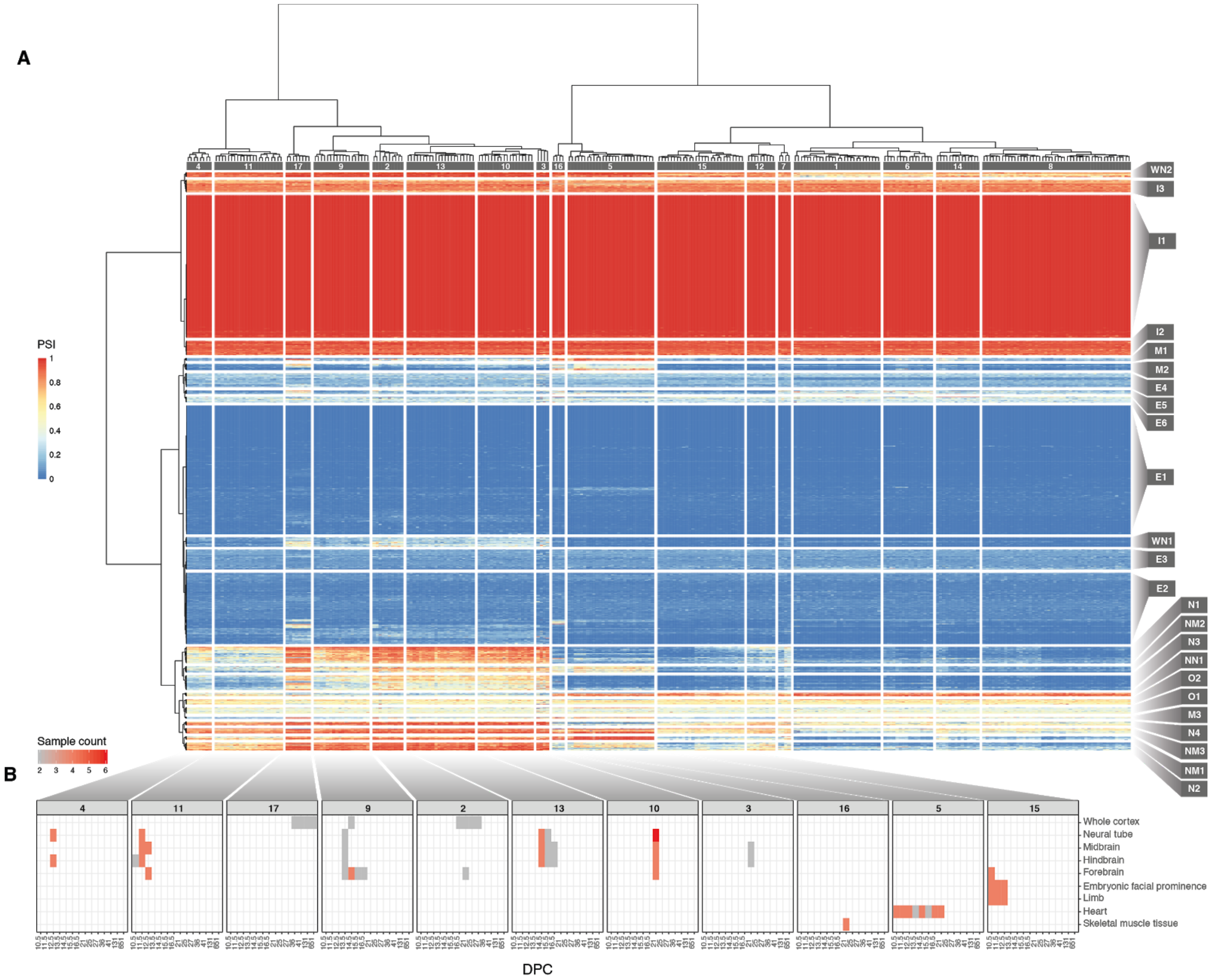

Figure 3.2. Microexon PSI biclustering A. Heatmap showing microexon inclusion patterns across analysed RNA-seq samples. Rows correspond to microexons and columns to RNA-seq samples. Blue to red colour scale represents PSI values. Microexon cluster names are shown on labels displayed at the right side. **B.** Tissue type and developmental stage composition from sample clusters containing neuronal samples or samples associated with high microexon inclusion.

The 24 microexon clusters were further analysed by dividing them into eight main categories based on the loading factors of the first two components from the PPCA (Figure 3.3). Assuming that PC1 and PC2 represent variance that can be associated with brain and muscle respectively, loading factors can be used as a proxy to evaluate the tissue-specificity behavior of microexon clusters. Following this logic, microexon clusters that have high mean loading factors (>0.03) for PC1 and PC2, were considered as neuromuscular (NM1-3). Clusters that have high loading for either PC1 or PC2 were considered as neuronal (N1-4) and muscular (M1-3), respectively. The remaining microexon clusters correspond to microexon that mostly have PSI inclusion levels that do not change across tissues. Thus, those microexon clusters that have an average PSI value lower than ⅓ were classified as Excluded (E1-6), while microexon clusters with a mean PSI value greater than ⅔ were classified as Included (I1-2). Only two clusters did not match any of the classification criteria mentioned above, so they were labeled as Other (O1-2). The number ID given to each microexon cluster corresponds to ranks computed based on PC1 or PC2 mean loading factors. By this way, N1 corresponds to the neuronal microexon cluster with the highest mean loading factor for PC1, while M1 is the muscular microexon cluster with highest mean loading factor for PC2.

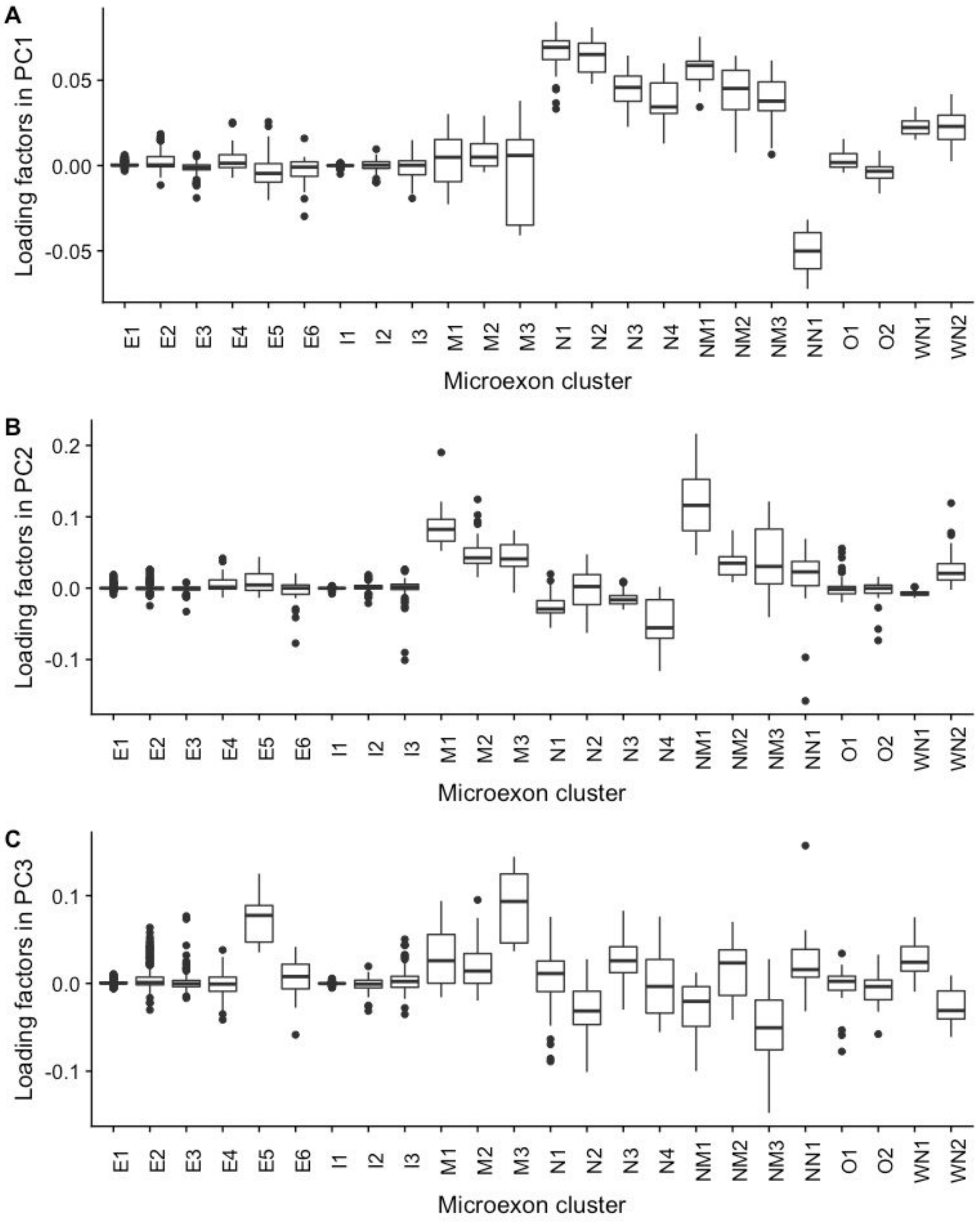

Figure 3.3: PPCA loading factors across microexon clusters. A-C. Letters in the x-axis denote the different microexon clusters.

Studies of standard alternative exons have shown that they typically have weaker splice signals than constitutive ones, and that they are less likely to disrupt the reading frame (Keren et al., 2010). Thus, I measured the splice site strengths as defined by the average U2 score of microexon flanking splice sites and the fraction of microexons that preserve the reading frame for each cluster (Fig 3.2d). As expected, the included clusters exhibit the strongest splicing signals, while the excluded clusters have the weakest splice sites, suggesting that constitutive inclusion of microexons relies on strong splicing signals. Moreover, the excluded clusters have a lower fraction of in-frame events, implying that they are likely to be more disruptive to gene function. Interestingly, neuronal, muscular and some neuromuscular clusters have almost as weak splice sites as the excluded clusters, but the fraction of in-frame events is on average 79.2%. This is considerably higher than the in-frame fractions for longer cassette exons (overall 43.2% and developmentally regulated 68.7%) (Weyn-Vanhentenryck et al., 2018). On the other hand, non-neuronal clusters have high U2 scores and also the highest in-frame microexon fraction. The in-frame fraction of each microexon cluster is strongly correlated with the conservation of the coding sequence (Pearson correlation = 0.86, p-value < 10^{-7}, Fig 3.3e), which implies that microexon clusters with higher conservation tend to preserve the protein frame.

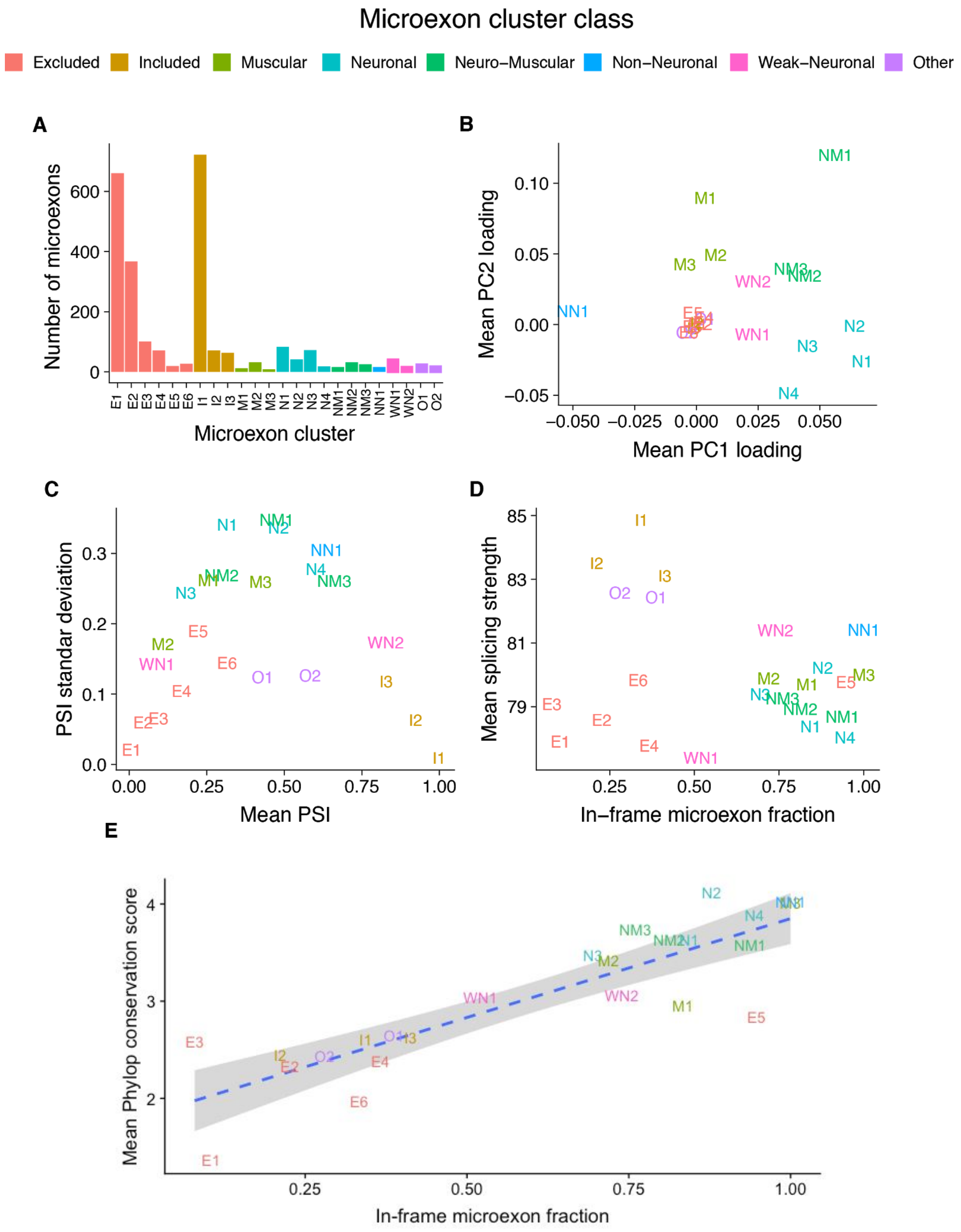

Figure 3.4: Inclusion properties of microexon clusters. **A.** Number of microexons belonging to each cluster. **B**. Mean loading factors across each cluster for PC1 and PC2. **C**. Mean and standard deviation of PSI values across microexon clusters. **D.** Mean U2 scores and in-frame fraction across microexon clusters. **E.** Relationship between genomic conservation and fraction of in-frame microexons for different microexon clusters.

Since microexons were previously shown to be progressively included during *in vitro* neuronal differentiation (Irimia et al., 2014), I hypothesized that neuronal and neuromuscular microexon clusters are progressively included throughout mouse embryonic brain development. Since PC1 strongly correlates with the developmental time from the samples (Fig 3.1), it can be considered as a proxy for early neuronal developmental time. To display how microexon PSI values relate to PC1, I calculated the average PSI value for microexon across tissue clusters and then I sorted these values according to the mean PC1 values of each tissue cluster (Fig 3.5). As expected, microexon clusters with higher mean PC1 loading factor values show greater mean PSI variability tissue clusters (Fig 3.3, Fig 3.5). Moreover, neuronal and neuromuscular clusters show a progresive increase of mean PSI inclusion values between tissue cluster number 11 and 2, which correspond to a range of tissue clusters that consist of neuronal samples extracted from increasingly older embryos (Fig 3.2, Fig 3.5). Moreover, across this same range of tissue clusters, non-neuronal microexons (NN1) show decreasing mean PSI values, which is in accordance with the negative loading factor values that were observed for this cluster. All these analyses suggest that there are groups of microexons that are progressively included at different rates during mouse embryonic development, while there is a minority group of microexon which follows the opposite trend.

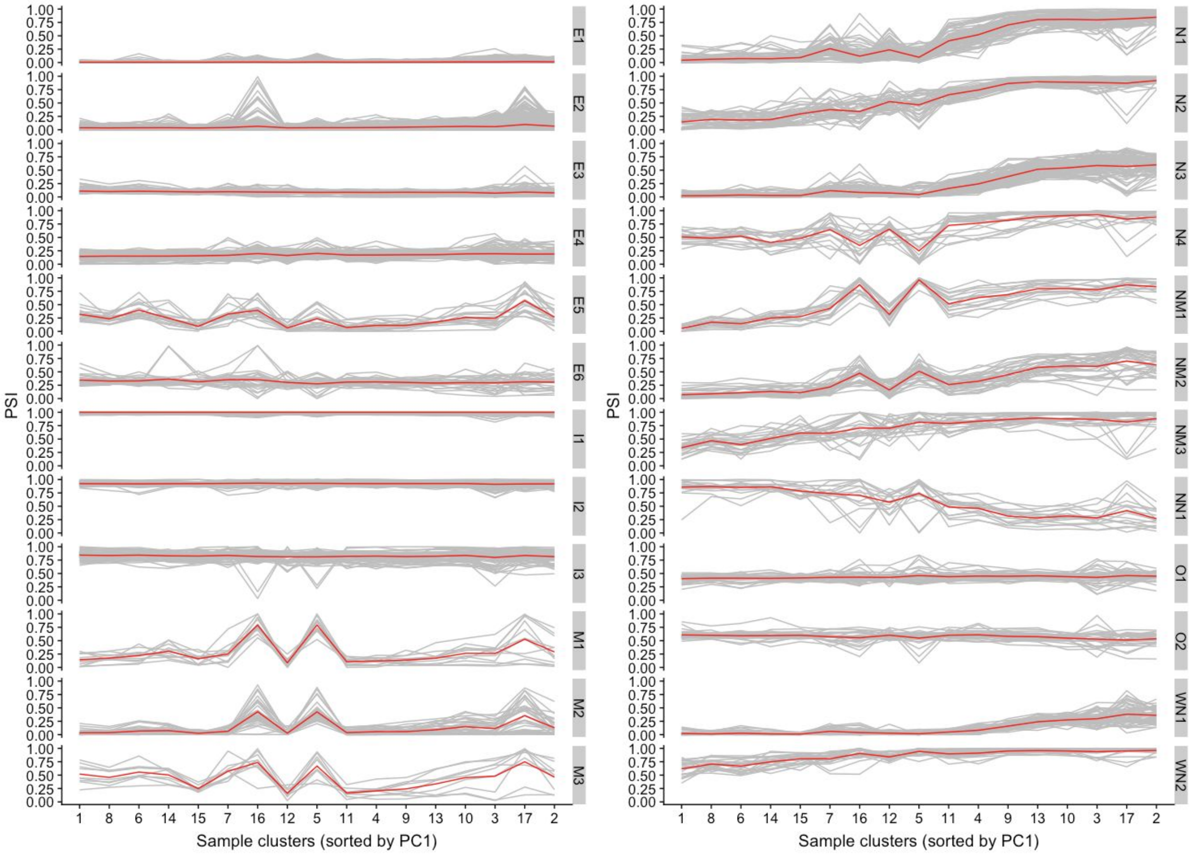

Figure 3.5: Microexon PSI values across all identified microexon clusters. Grey lines correspond to individual microexons, while red lines denote the average PSI value for a given microexon cluster across sample clusters.

In order to compare the PSI variation across mouse embryonic brain development, I defined the group of tissue clusters with lowest absolute values of mean PC1 loading factors (C1, C6 and C8) as baseline for null neuronal microexon inclusion (negative control). As expected, the contrast of the mean PSI values between these values and neuronal and muscular samples revealed distinct patterns across neuronal, neuro-muscular and non-neuronal clusters (Fig 3.6).

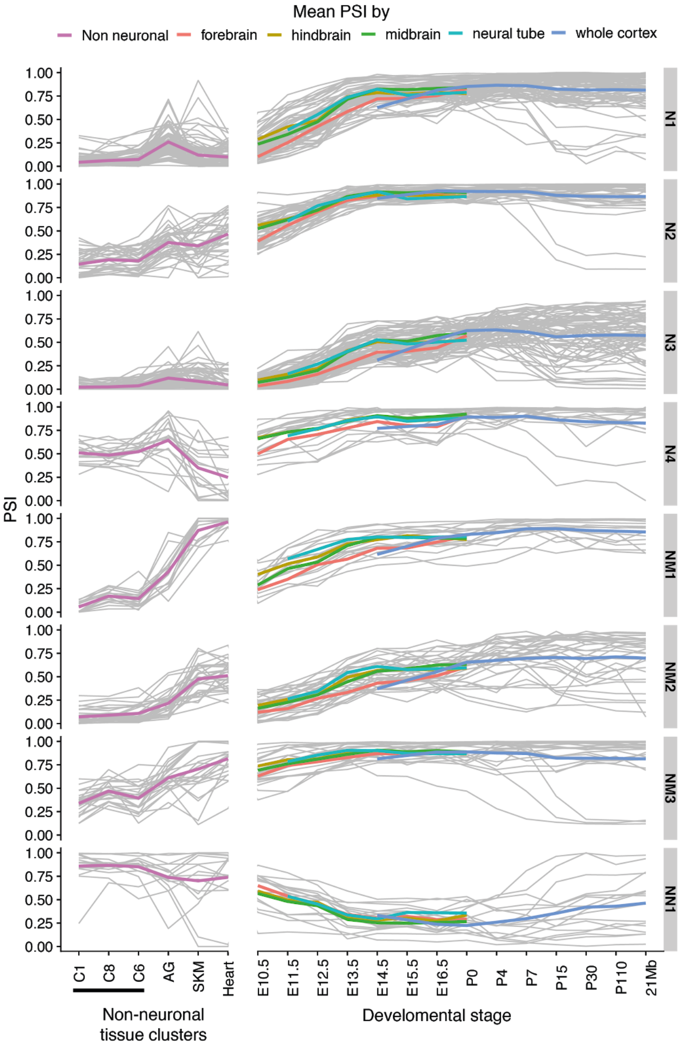

Figure 3.6: Mean PSI values across neuronal and neuromuscular microexons. Each grey line represents mean PSI values for a microexon across all samples from a tissue cluster or neuronal developmental stage (x-axis).

To quantitatively assess alternative splicing across different sample sets, I integrated Whippet (Sterne-Weiler et al., 2018), which provides a module for quantifying splicing events (whippet-quant) and a statistical framework to assess alternative splicing events (whippet-delta). Given an input gene annotation file, Whippet builds contiguous splice graphs (CSGs) to represent each transcript. In a CSG nodes represent non-overlapping exonic sequences, while edges represent splice junctions or contiguous exonic regions (Fig 3.7). Since the reads are directly mapped to the CSG, Whippet enables a fast annotation-oriented quantification of splicing events. Thus, I integrated Whippet as an optional microexon re-quantification module downstream of the MicroExonator discovery module. For this purpose, MicroExonator integrates the final list of high confidence microexons into the gene annotation file and generates a gene transfer file (GTF) which enables Whippet to quantify annotated and novel microexons, in addition to other alternative splicing events. To incorporate MicroExonator quantification results into Whippet's statistical framework, PSI values for microexon splicing nodes are replaced by the ones obtained by MicroExonator. In a later step, both Whippet and MicroExonator based quantifications are used to assess alternative microexon inclusion across the given set of comparisons.

[redacted]

Figure 3.7: An overview of Whippet's computational workflow to quantify alternative splicing events. A. Illustration of Whippet's Node assignment given an example gene annotation with two isoforms. **B.** Representation of the CSG model that would be built given the example gene annotation provided above. **C.** Transcriptome indexing from CSGs generated for each annotated gene. **D.** Read alignment to the indexed transcriptome. E. Alternative splicing quantification through node PSI estimation, which takes into account the full set of RNA-seq reads aligned to edges that connect or exclude the corresponding splicing nodes. This figure was taken from Sterne-Weiler *et. al* 2018.

The implementation of the Whippet quantification module enabled the systematic assessment of microexon alternative splicing events across mouse embryonic brain development. RNA-seq samples from midbrain, hindbrain and neural tube (MHN) were grouped by their correspondent developmental stage and compared with the previously defined negative control using whippet-delta. The evaluation of microexon alternative splicing events detected using both Whippet and MicroExonator, shows an increasing number of inclusion events throughout mouse embryonic brain development (Fig 3.8a-b), which is consistent with the gradual inclusion of neuronal microexons observed in Fig 3.6. High correlation values can be observed between delta PSI values for microexon splicing nodes quantified with Whippet and MicroExonator (Fig 3.8c). However, correlation values obtained across different microexon splicing node types differ substantially (Fig 3.8d). While most microexon microexons splicing nodes are CE type (here referred as mCE) and flanked by strictly intronic regions, some CE are also flanked by AA or AD splicing nodes that represent the inclusion of a longer microexon (mAA or mAD) or an exon longer than 30 nt (AA or AD). Correlation between delta PSI values calculated using Whippet and MicroExonator is highest for mCE, mAA and mAD splicing nodes (Fig 3.8d). By contrast, microexons that are flanked by exonic (CE_mAA / CE_mAD) or microexonic splicing nodes (mCE_mAA / mCE_mAD) had significantly lower correlation. These splicing nodes are frequently derived from complex alternative splicing events where microexons could be completely skipped or included in a shorter form. Whippet was reported to perform particularly well for complex alternative splicing events (Sterne-Weiler et al., 2018). However, Whippet PSI measurements for mCE_mAA and mCE_mAD splicing nodes are highly correlated with their corresponding mAA and mAD nodes (Fig 3.9e), suggesting that their measurements may not be independent under Whippet quantification model. Some of these highly correlated microexon pairs exhibit lower correlations when quantified by MicroExonator, suggesting active competition between shorter and longer microexons. Since the MicroExonator quantification module is only based on the relative number of spliced reads that represent each set of splicing paths that are compatible or incompatible with microexon inclusion, it was able to disentangle the inclusion on microexon associated to alternative 5′/3′ splice sites. Competition of short and longer forms of microexons have already been reported to have a key role for LAR-RTP protein function in synaptic adhesion, thus a precise quantification of these events might contribute to deeper understanding of neuronal microexon splicing (Won and Kim, 2018; Yamagata et al., 2015a).

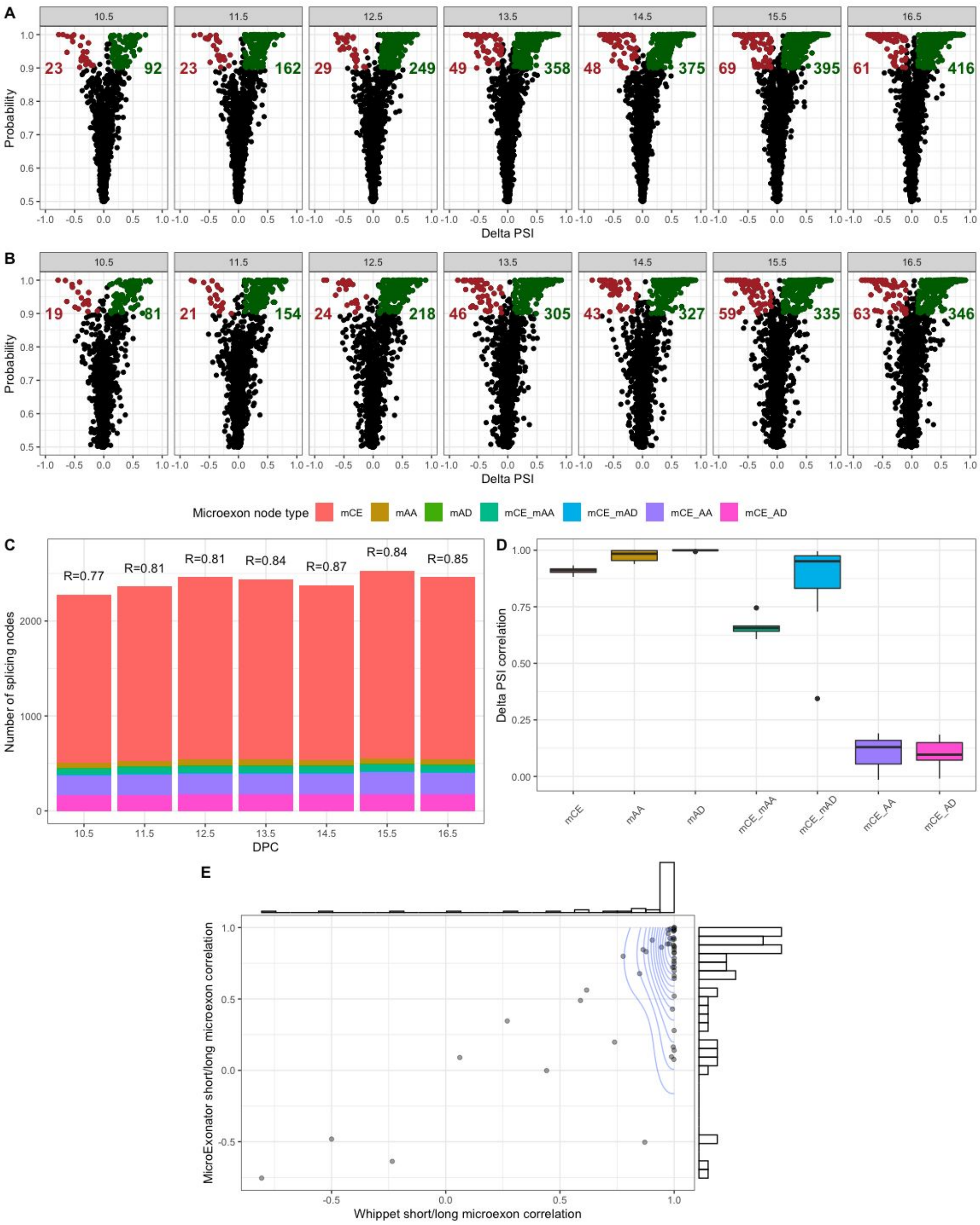

Figure 3.8: Differential inclusion analysis performed MicroExonator and Whippet quantification outputs show similar trends. A-B. Volcano plots showing the distribution of delta PSI values of microexon splicing nodes and their corresponding probability of being differentially included across MHN samples coming from different developmental stages (E10.5-E16.5). Delta PSI measurements were calculated using Whippet (A) or MicroExonator (B) microexon inclusion quantification. Alternatively included splicing nodes are highlighted in red (excluded) and green (included). Coloured numbers indicate the corresponding quantity of each group of differentially included splicing nodes. **C.** Abundance of splicing nodes quantified across the different comparisons classified according to the different classes mentioned above. Number on top indicate Perason’s correlation index values (R). **D.** Correlation between mCE_AA / mCE_AD and their flanking mAA / mAD splicing nodes on Whippet and MicroExonator quantification.

I found 422 microexons that were consistently detected as differentially included on both Whippet and MicroExonator splice node quantification across at least one of the MHN comparisons performed against the defined base group. Interestingly, 323 of these microexon changes are maintained for all subsequent stages once they have been observed, meaning that they correspond to stable transcriptome signatures that are acquired during embryonic mouse brain development (Appendix - Table I). The distribution of the developmental stages when these sustained microexon changes started to be detected differed. While some microexon clusters showed early changes (N1 and N2), other clusters started to be differentially included later on (N3, NM1 and NM2) (Fig 3.9a). As forebrain tends to show delayed microexon inclusion compared to midbrain, hindbrain and neural tube (Fig 3.1c, 3.6), I pooled forebrain samples between E10.5 and postnatal (P0) and compared samples grouped by developmental stage with the non-neuronal control sample group. I found 401 microexons that were differentially included during at least one forebrain developmental stage, with 258 that were sustained through all later developmental stages (Fig 3.9b). While all the observed microexon changes across neuronal and neuromuscular clusters correspond to inclusion events, microexons from the non-neuronal cluster (NN1) only correspond to exclusion (Fig 3.9a-b).

In agreement with previous studies (Irimia et al., 2014; Li et al., 2015) I also found strong inclusion patterns associated with heart and SKM. In addition, I found microexon inclusion patterns associated with AG samples (Fig 3.1a-b, 3.6). Compared with the set of non-neuronal control samples, I found 81, 109 and 58 microexons to be differentially included in heart, SKM and AG respectively (Fig 3.9c). Most neuronal and neuromuscular microexon clusters show distinct microexon inclusion patterns compared to controls, whereas non-neuronal clusters were associated with microexon inclusion events in heart or exclusion events in SKM samples (Fig 3.9c).

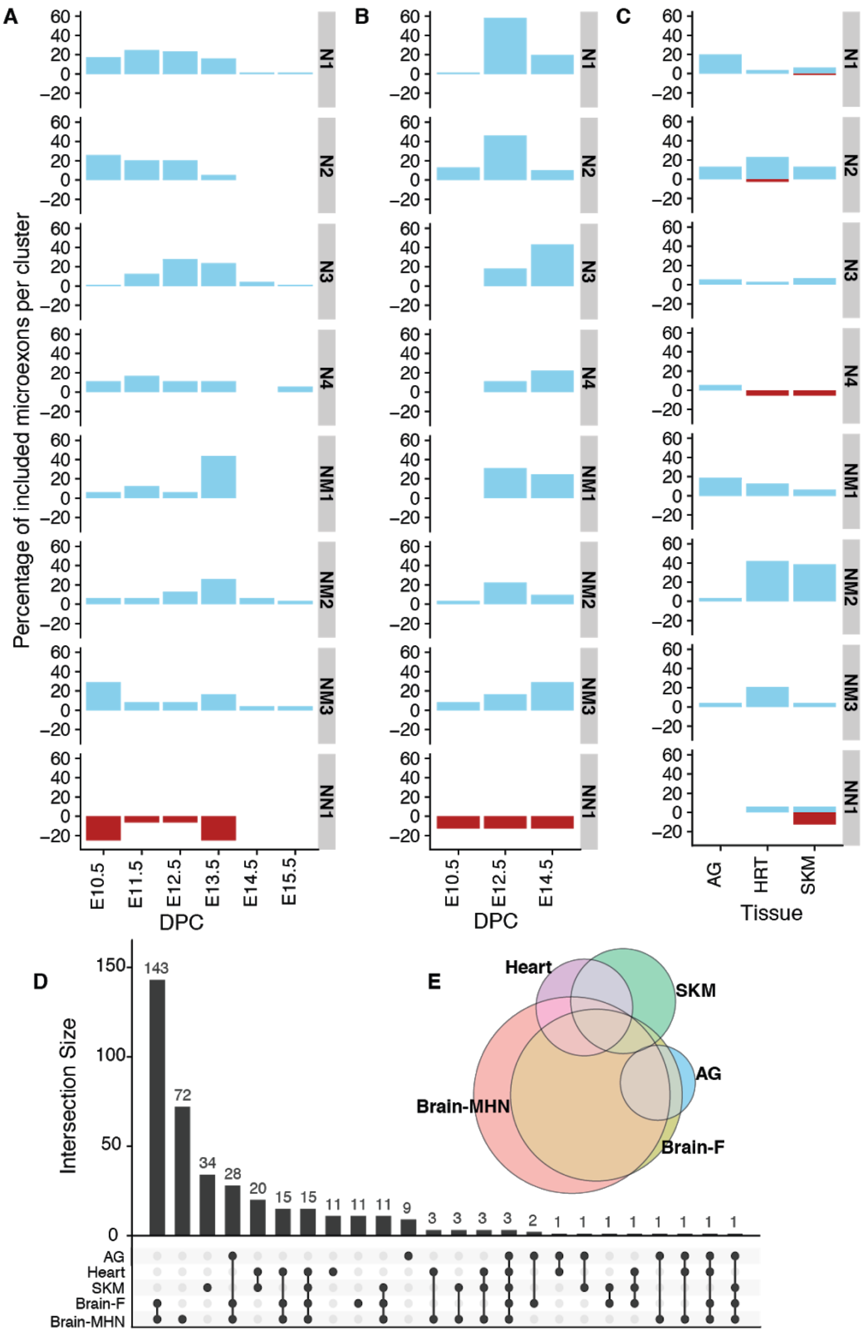

Figure 3.9: Differential inclusion analysis of microexons. A-C. Alternative microexons detected between non-neuronal tissue samples and midbrain, hindbrain and neural tube (F); forebrain (G); adrenal gland (AG), heart (HRT) and skeletal muscle (SKM) (H). Microexon splicing changes are represented as the percentage of microexons corresponding to each microexon cluster, where microexon inclusion fractions are represented with blue bars and exclusion events with upside down red bars. **D.** Intersection between microexon sets that were differentially included across sample groups. The vertical bars show the number of microexons corresponding to combinations indicated by the connected dots below. **E.** Area-proportional Euler diagram representing the most abundant intersections between differentially included microexon sets.

The set of microexons that were differentially included across the different tissue groups (brain-MHN, forebrain, heart, SKM and AG) overlap. Closer inspection reveals high concordance between the set of microexons associated with sustained changes in inclusion across MHN and forebrain samples. Surprisingly, I found a significant overlap of alternatively included microexons that have concordant patterns in AG and neuronal samples (hypergeometric test p-value < 10^{-30}). Nearly all of the AG microexons are also found in neuronal samples (Fig 3.9d-e), but in AG I observed lower PSI values (Fig 3.10). I hypothesize that the mixture between neuronal and non-neuronal isoforms found in AG is due to the chromaffin cells in the adrenal medulla which are derived from the neural crest and share fundamental properties with neurons (Bornstein et al., 2012; Shtukmaster et al., 2013).

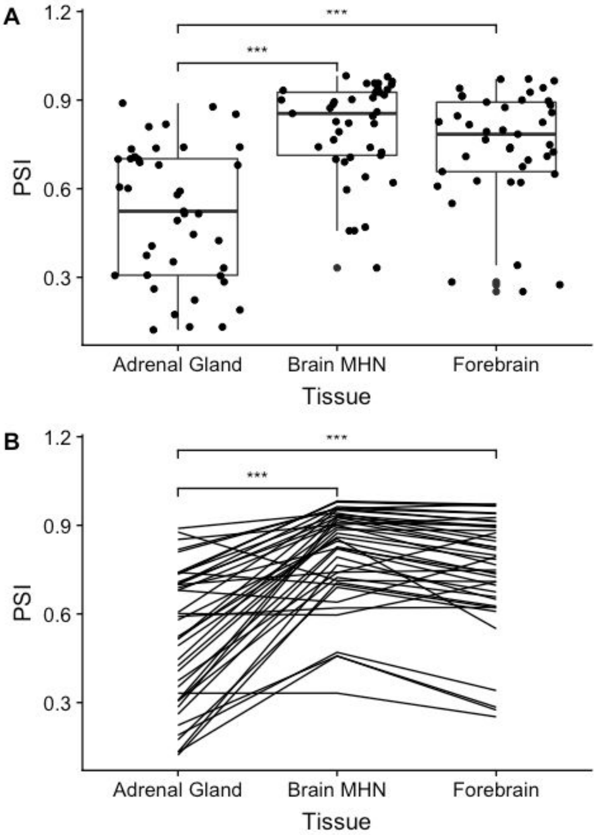

Figure 3.10: Differences in PSI score between adrenal gland, brain MHN and forebrain tissues. A. Shown by box-plots superimposed with jittered dot-plots. **B.** Shown by line-plots. Statistical differences were assessed by Wilcoxon test while correcting for multiple comparisons. Significant p-values are denoted by * (>0.05), ** (>0.01) and *** (>0.001).

3.2.2 Microexon alternative splicing is coordinated throughout embryonic development

Based on *in vitro* studies of neuronal differentiation, it has been proposed that microexons are an integral part of a highly conserved alternative splicing network (Irimia et al., 2014). Our analysis of mouse embryonic data (Fig 3.6) shows that most microexons remain included once their splicing status has changed. To explore possible functional consequences of these splicing changes I analyzed the interactions between the proteins which contain microexons by constructing tissue specific protein-protein interaction (PPI) networks for brain, heart, SKM and AG using STRING (Szklarczyk et al., 2017). For all four PPI networks the degree of connectivity was significantly higher than expected by chance given the same number of nodes (p-value<10^{-16}). On average, there were 2.7-fold more connections than expected by chance, with brain having the largest number of connections (Appendix - Table II). Next, I considered the gene ontology (GO) terms and pathways associated with the PPI networks (Fabregat et al., 2018). The Reactome pathways that showed a significant enrichment, include parts of molecular complexes that are involved in membrane trafficking pathways, e.g. “ER to Golgi anterograde transport”, “Clathrin-mediated endocytosis”, “Golgi associated vesicle biogenesis”, “Intra-Golgi and retrograde Golgi-to-ER traffic” and “Lysosome vesicle biogenesis” (Fig 3.11a-b). I also found a distinct cluster that is annotated as part of “Protein-protein interactions at synapses" (Fig 3.11d). This group includes presynaptic proteins, e.g. liprins (*PPFIA1*, *PPFIA2* and *PPFIA4*), protein tyrosine phosphatase receptors (*PTPRF*, *PTPRD* and *PTPRS*) and neurexins (*NRXN1* and *NRXN3*), which are involved in trans-synaptic interactions with multiple postsynaptic proteins, having a key role in synaptic adhesion and synapse organization. The interactions of these proteins have been shown to be highly regulated by alternative splicing (Takahashi and Craig, 2013), and our results reveal that many of these events occur towards the end of embryonic development (Fig 3.11f).

In agreement with previous reports that have highlighted the importance of microexons for axonal and neurite outgrowth (Ohnishi et al., 2017; Quesnel-Vallières et al., 2015), I detected 18 proteins in the PPI network that are annotated as part of the "Axon guidance" Reactome pathway. These proteins are found in the center of the network and they are connected with the domains involved with membrane trafficking and transsynaptic protein-protein interactions (Fig 3.11a-e). For two of the proteins associated with this pathway, the non-receptor tyrosine kinase protein SRC and L1 cell adhesion molecule (L1cam), microexon inclusion is known to play a key role in neuritogenesis (Kamiguchi and Lemmon, 1998; Keenan et al., 2017), but the importance of microexons in other proteins in this pathway remains poorly characterised. At early developmental stages (E10.5-E11.5) I found several microexon alternative splicing events in genes associated with "membrane trafficking" pathways concentrated. A subset, "clathrin mediated endocytosis" is associated with microexon changes in the later stages, as most events became significant only after E12.5 (Fig 3.11g). Similarly, "axon guidance" microexon changes mostly occur at E11.5, in particular the microexon alternative splicing events for proteins that interact with L1cam.

Since microexon inclusion occurs in several waves (Fig 3.6), I hypothesized that the temporal dynamics would be reflected in the topology of the PPI network. To quantitatively evaluate the position of each gene in the network, I calculated several centrality measures. The result is not straightforward to interpret since several of the central nodes feature more than one microexon inclusion event (e.g. *SYNJ1*, *ANK3* and *DCTN2*), which sometimes emerge at different embryonic stages . Nevertheless, the results show that L1cam and 6 out of 10 of its interactors are amongst the 15% of nodes with highest eigencentraly and that *SRC* has the highest harmonic centrality and betweenness. An investigation of genes corresponding to some of the most relevant GO terms revealed that proteins located at more central positions of the network (measured as eigencentrality[7]), have microexons that are included

[7] Eigencentrality, also known as eigenvector centrally, is a measure of node centrality that is computed based on the eigenvectors of the adjacency matrix. This method assigns higher centrality to nodes that are more connected, particularly to those that are also connected with other highly central nodes.

earlier in mouse embryonic brain development (Kruskal-Wallis rank sum, p-values < 0.05) (Fig 3.11h-i).

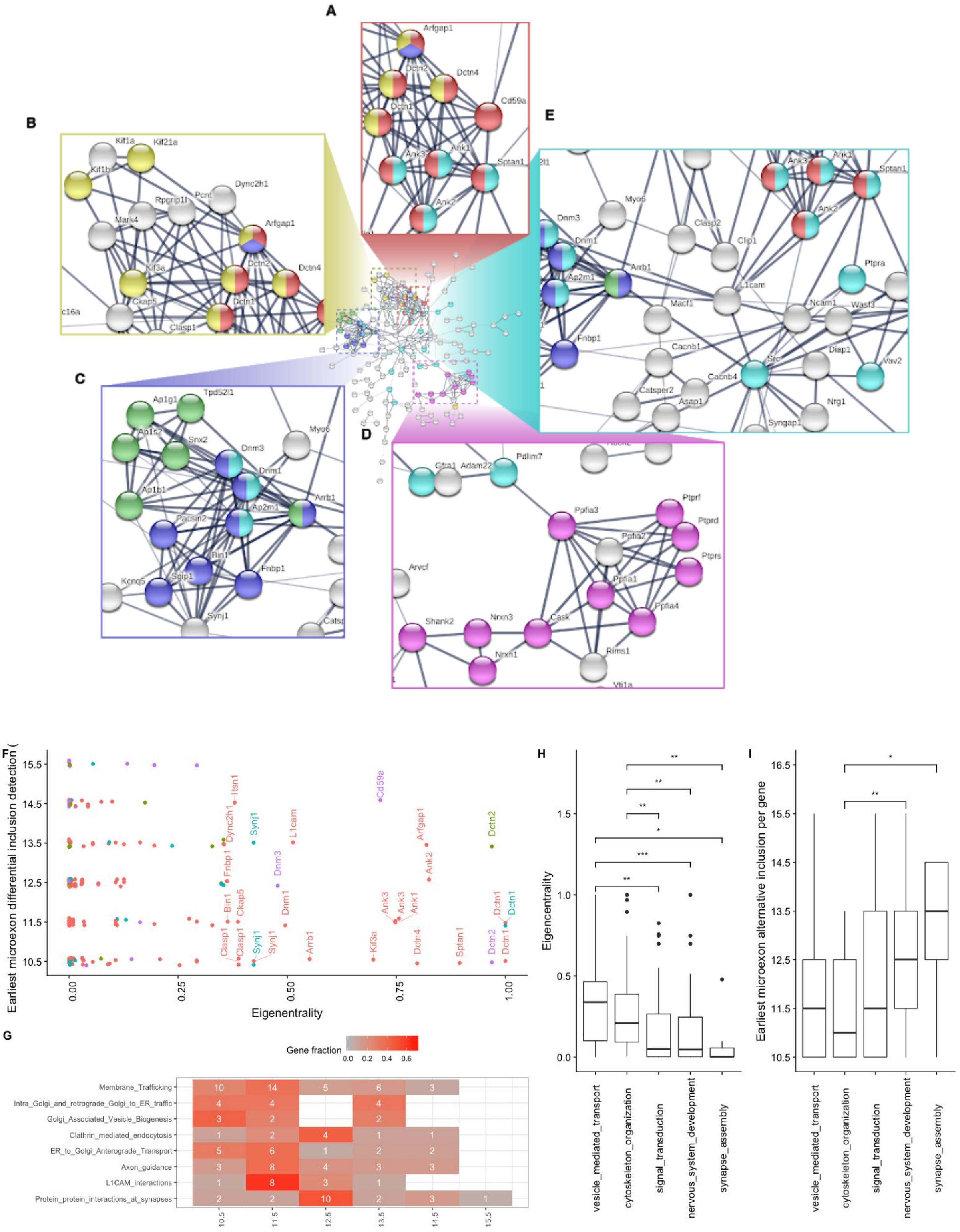

Figure 3.11: Microexon protein-protein interaction network. A-E) PPI network using as input genes that have microexons that are differentially included across mouse embryonic brain development. Colours represent different Reactome pathways that were enriched on the network; Axon guidance (light blue), Protein-protein interactions at synapses (pink), ER to Golgi anterograde transport (red), Clathrin-mediated endocytosis (dark blue), Golgi associated vesicle biogenesis (green), Intra-Golgi and retrograde Golgi-to-ER traffic (yellow). F) Eigencentrality calculated for each gene node in relation to the developmental stage at which each microexon was included. G) Effect of microexon alternative splicing over different Reactome pathways. Counts indicate the number of microexons that start to be differentially included at each developmental stage for different Reactome pathways that were significant after taking the whole genome as background. H-I) Eigencentrality and earliest developmental stage at which each gene is affected by differential microexon inclusion show differences across some of the GO categories that were significantly enriched after gene background correction. Statistical differences were assessed by Wilcoxon test while correcting for multiple comparisons. Significant p-values are denoted by * (>0.05), ** (>0.01) and *** (>0.001).

3.2.3 MicroExonator enables the identification of novel neuronal microexons

Of the 343 microexons that were differentially included across brain development, 90 were not consistently annotated between GENCODE and VastDB. I found 26 neuronal microexons that are only annotated in GENCODE, and 33 neuronal microexons that are not annotated in GENCODE, but are present in VastDB. Despite the fact that the mouse genome is comprehensively annotated, I found 35 neuronal microexons that are not annotated in GENCODE nor VastDB. Due to the high sensitivity and specificity demonstrated in simulations (Fig 2.4), I expect that all 34 microexons >4 nts are true positives.

To validate one of the novel microexons, I focused on the Dctn2 gene (eigencentrality of 0.76), where I detected two adjacent differentially included microexons of length 9 and 6 nts (Fig 3.4a). Neither of these microexons are annotated in GENCODE, but the 9-nt microexon are annotated in VastDB (MmuEX0013953). Interestingly, the downstream 6-nt microexon that was discovered by MicroExonator is validated by spliced ESTs (Benson et al., 2004). I detected differential inclusion of the 6-nt Dctn2 microexon from E10.5 in MHN samples, whereas in forebrain it is differentially included from E12.5 (Fig 3.4b).

Hugo Fernandez performed qRT-PCR experiments to assess the inclusion of the Dctn2 6-nt microexon during a mESC to neuron differentiation protocol using one set of primers that were designed to amplify Dctn2 isoforms with 6-nt microexon inclusion and another set to amplify total Dctn2 isoforms. After normalizing the qRT-PCR values using dilution series of neuronal samples, I calculated the ratio of 6-nt inclusion across mESC, EPI cells and differentiated neurons at two different stages (Fig 3.4c). The inclusion ratios from the qRT-PCR measurements indicate that the Dctn2 6-nt microexon is included through *in vitro* differentiation of mESC to neuron, consistent with our findings during embryonic development for this microexon. These results show that the alternative splicing quantification provided by MicroExonator can identify novel microexons, even for model organisms that are well annotated.

3.2.4 Identification of microexons in zebrafish brain.

To demonstrate how MicroExonator can be applied to species with less complete annotation, I analyzed 23 RNA-seq samples from zebrafish brain (Park and Belden, 2018). I found 1,882 microexons, of which 23.8% are not found in the ENSEMBL gene annotation. I used liftover (Hinrichs et al., 2006) to assess whether some of these microexons are evolutionarily conserved microexons in mouse, and I successfully mapped 401 zebrafish microexons. Of these, 85% mapped directly to a previously identified mouse microexon, and most of the remaining 15% mapped to longer exons. Mapping the microexons in the other direction, 617 out of the 2,938 that were identified from the mouse development data mapped to the zebrafish genome and 49.7% of those in return mapped to a zebrafish microexon. By integrating these results I obtained a total of 402 microexon pairs that are found in both zebrafish and mouse. Since 90.3% of the pairs had identical length in both species, they are highly likely to correspond to evolutionarily conserved microexons. I calculated the percentage of conserved microexons between mouse and zebrafish for each mouse microexon cluster, and I found that microexon clusters involved in neuronal regulation (Neuronal, Neuro-muscular, Non-Neuronal and Weak-Neuronal clusters) have a significantly higher degree of conservation than the other microexon clusters (two-sided Wilcoxon test, p-value < 0.01, Fig 3.4d).

To compare the microexon annotation between mouse and zebrafish, I calculated the number of conserved microexons between these two species that are missing in mouse or zebrafish gene transcript annotation. While only 6.9% of these exons are missing from the mouse transcript annotation provided by GENCODE, 16.1% are missing from the ENSEMBL zebrafish transcript annotation. Moreover, the largest fraction of conserved microexons that are missing in zebrafish transcript annotation corresponds to neuronal microexons (Fig 3.4e).

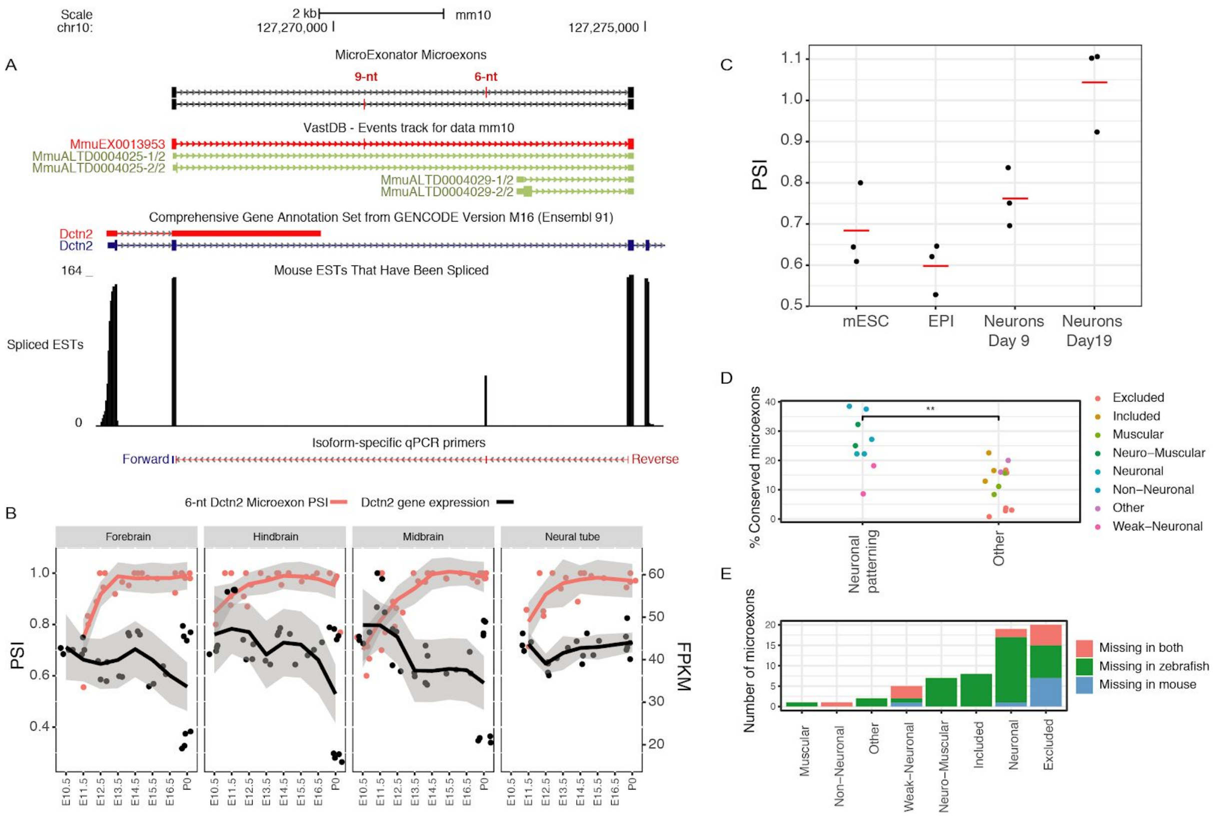

Figure 3.12: Discovery of novel microexons in mouse and zebrafish. A. Alternative Dctn2 microexons that are inconsistently annotated in mouse GENCODE and VastDB annotations. **B.** Novel 6-nt Dctn2 microexon shows a progresive inclusion through mouse embryonic development. **C.** PSI values calculated from normalized RT-PCR measurements show a gradual inclusion of the 6-nt Dctn2 microexon though *in vitro* neuronal mESC to neuron differentiation. **D.** Microexon clusters that exhibit neuronal patterning have higher conservation percentage between mouse and zebrafish than the other microexon clusters. Every dot

corresponds to a different microexon cluster and the colour indicates its type. ** denotes p-value<0.01 calculated for a two-sided Wilcoxon test. **E.** Number of conserved microexons between mouse and zebrafish that are missing from their transcript annotation.

3.2.5 Cell type specific microexon inclusion in mouse visual cortex.

Our analysis of neuronal development suggested that the main difference in microexon inclusion is between time points rather than tissues. However, these data do not reflect the diversity of cell types within neuronal tissues, and since the neural cortex is one of the most diverse tissues in the murine body, I hypothesized that microexon inclusion patterns may vary amongst different subcellular types that can be found in the adult mouse neuronal cortex. Full length scRNA-seq experiments using the SMART-seq2 protocol have enabled the identification of two main neuronal classes, glutamatergic and GABA-ergic neurons, and seven non-neuronal cell-types (Tasic et al., 2016). Despite the 3′ bias previously reported for SMART-seq2 protocol, these scRNA-seq experiments enabled Tasic and co-workers to evaluate exon usage and identify alternative splicing events across cell-types. Thus, I developed a downstream module of MicroExonator to perform alternative splicing analysis of microexons using full-length single-cell data. I used it to identify microexon alternative splicing events between GABA-ergic and glutamatergic neurons defined by Tasic et al., 2016, containing 739 and 764 cells, respectively.

I first ran the microexon discovery module with an expanded annotation, which included the microexons discovered from our previous analyses. This yielded 2,344 microexons that were included in at least one cell. Next, I used Whippet to quantify the PSI of the microexons detected by MicroExonator for each cell. Since alternative splicing analysis heavily relies on the number of splice junction reads detected, the sparsity of read coverage scRNA-seq is a technical challenge that needs to be overcome in order to reliably identify alternative splicing events. Thus, for each neuronal type I systematically pooled GABA-ergic or glutamatergic neurons into pseudo-bulk groups of 15 cells, which were subsequently quantified by Whippet using an indexed transcriptome that considers all the novel microexons identified during the analysis. The analysis of pseudo bulk PSI values identified a total of 39 differentially included microexons, 20 of which were also identified from the single cell PSI values (Fig 3.13). Moreover, all of these steps were implemented as an optional extension of the core snakemake workflow of MicroExonator, which means it can be used to identify alternative splicing events between other groups of cells profiled using full length scRNA-seq protocols.

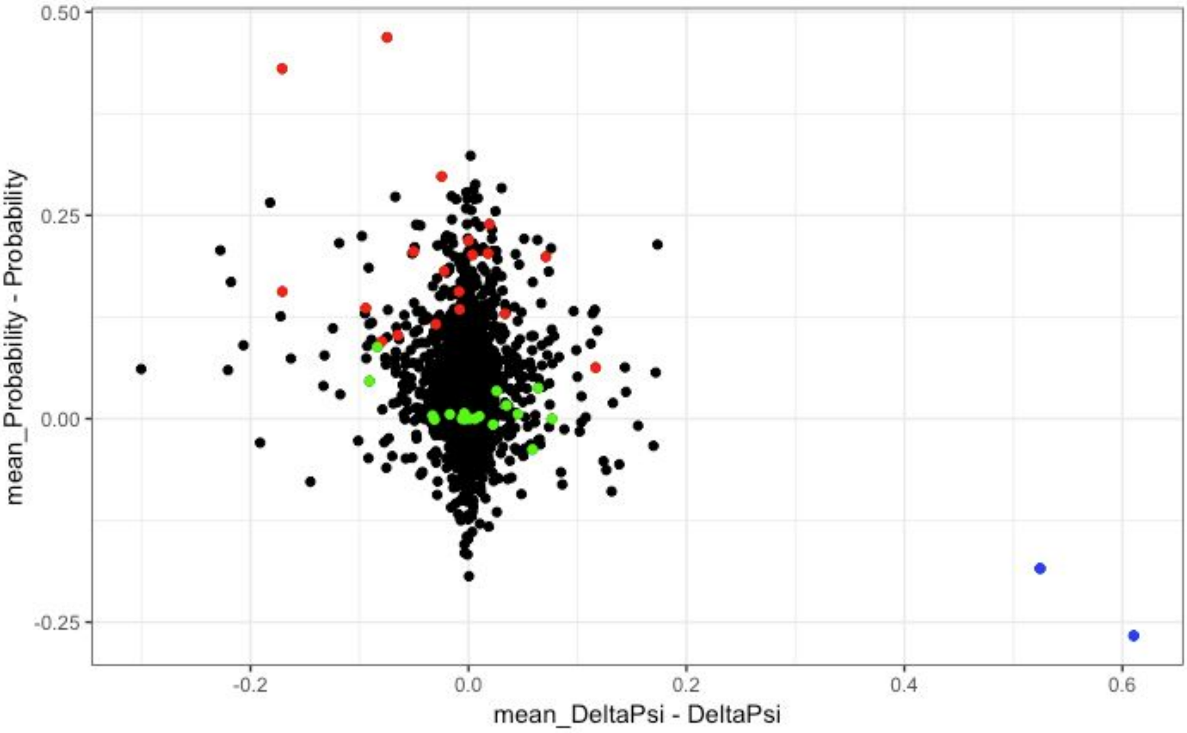

Figure 3.13: Differences between unpooled and pooled methodologies to assess microexon splicing changes in single cell data. For each microexon, different delta PSI and probability values were obtained while GABAergic and glutamatergic data were processed through pseudo pooling (pooled) or standard analysis (unpooled). To get an idea of the consistency of the results across these two methodologies. Since the pseudo pooling process was done ten times to avoid random arrangement effects, the comparison between pooled and unpooled strategies can be measured as the mean of the pooled results (delta PSI and probability) minus the ones obtained by the unpooled approach.

Among the genes that contain differentially included microexons between GABA-ergic and glutamatergic neurons is a group of eleven genes that encode for proteins that localize at synaptic compartments. I found seven presynaptic proteins, two postsynaptic proteins and two proteins that have been observed at both locations (Fig 3.14a). For example, the type IIa RPTPs subfamily of proteins undergoes tissue-specific alternative splicing that determines the inclusion of four short-peptide inserts, known as mini-exon peptides (meA-meD) (Pulido et al., 1995a, 1995b; Takahashi and Craig, 2013). While meB comprises four residues (ELRE) and is encoded by a single microexon, meA has three possible variants that can form as a result of the combinatorial inclusion of two microexons; meA3 (ESI), meA6 (GGTPIR) and meA9 (ESIGGTPIR) (Yamagata et al., 2015a). Our analysis shows a consistent inclusion of meB in both GABA-ergic and glutamatergic neurons. However, I detected cell type specific rearrangement of meA microexons which promotes inclusion of meA9 in glutamatergic neurons, while in GABA-ergic neurons meA variants are mostly excluded (Fig 3.14b). Alternative splicing of meA/B microexons are key to determining the selective trans-synaptic binding of PTPδ to postsynaptic proteins, which is a major determinant of synaptic organization (Takahashi and Craig, 2013). In addition, I found other alternatively spliced microexons in genes that are involved in synaptic cell-adhesion, e.g. Gabrg2, Nrxn1 and Nrxn3 (Südhof, 2017; Takahashi and Craig, 2013). The microexon inclusion in these genes is variable across the core clusters, sometimes showing stark differences between GABA-ergic and glutamatergic neuron subtypes (Fig 3.15). These results suggest that microexon inclusion is not only coordinated at the tissue-type level, but that it is also finely tuned across neuronal cell-types, and these differences may be of importance for determining neuronal identity.

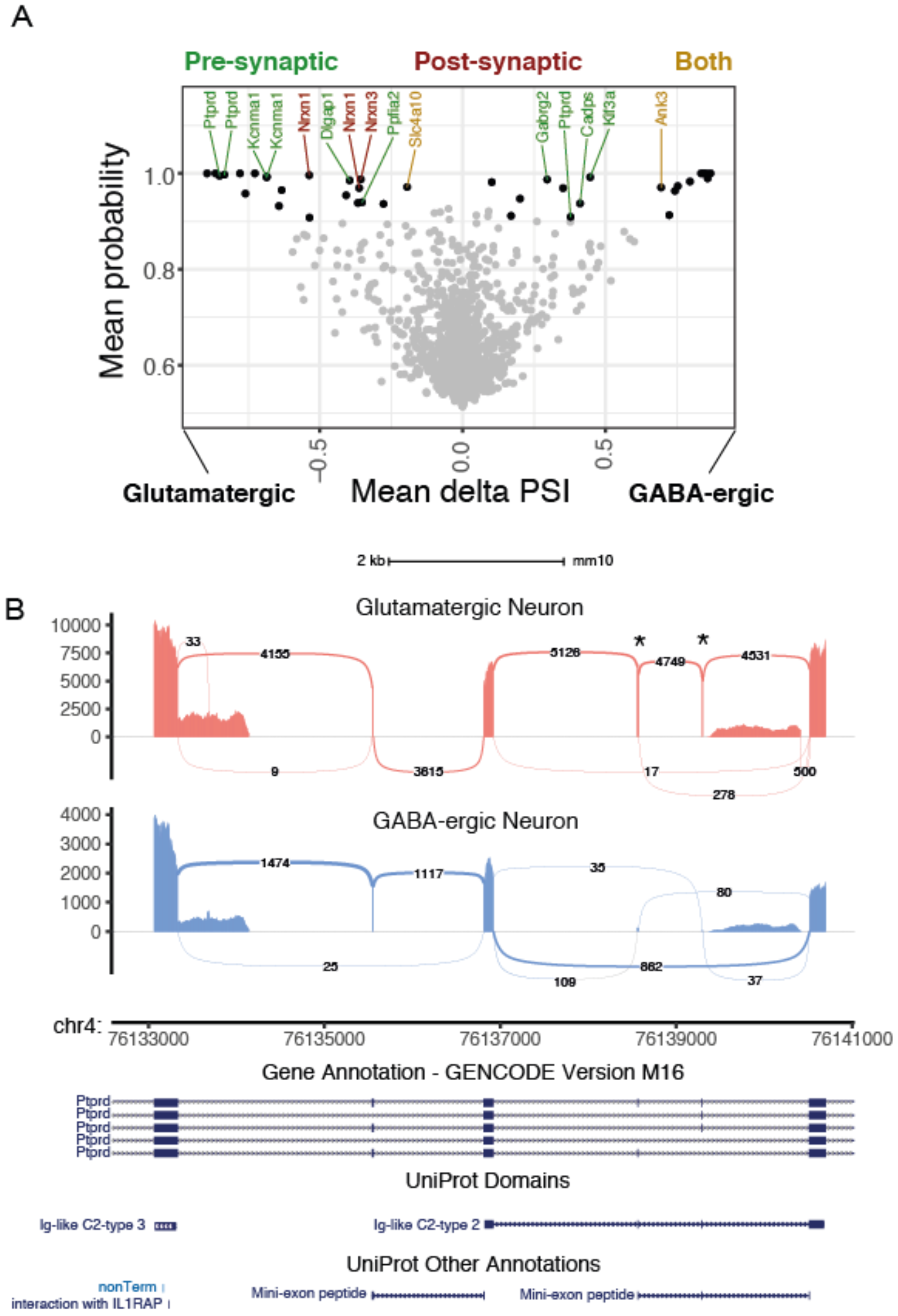

Figure 3.14: Differential alternative splicing analysis of microexons between glutamatergic and GABA-ergic neurons. A. Volcano plot showing an overview of the alternatively included microexons between glutamatergic and GABA-ergic neurons. Differentially included microexons are highlighted in black. Detected synaptic proteins containing cell-type specific microexons are labeled with different colours depending on their sub-synaptic localization. **B.** Sashimi plot showing PTPδ microexons that determine the inclusion of meA/B mini-exon peptides. Numbers indicate the amount of splice reads that support each splice junction and * denote microexons that were detected as differentially included between glutamatergic and GABA-erigic neurons.

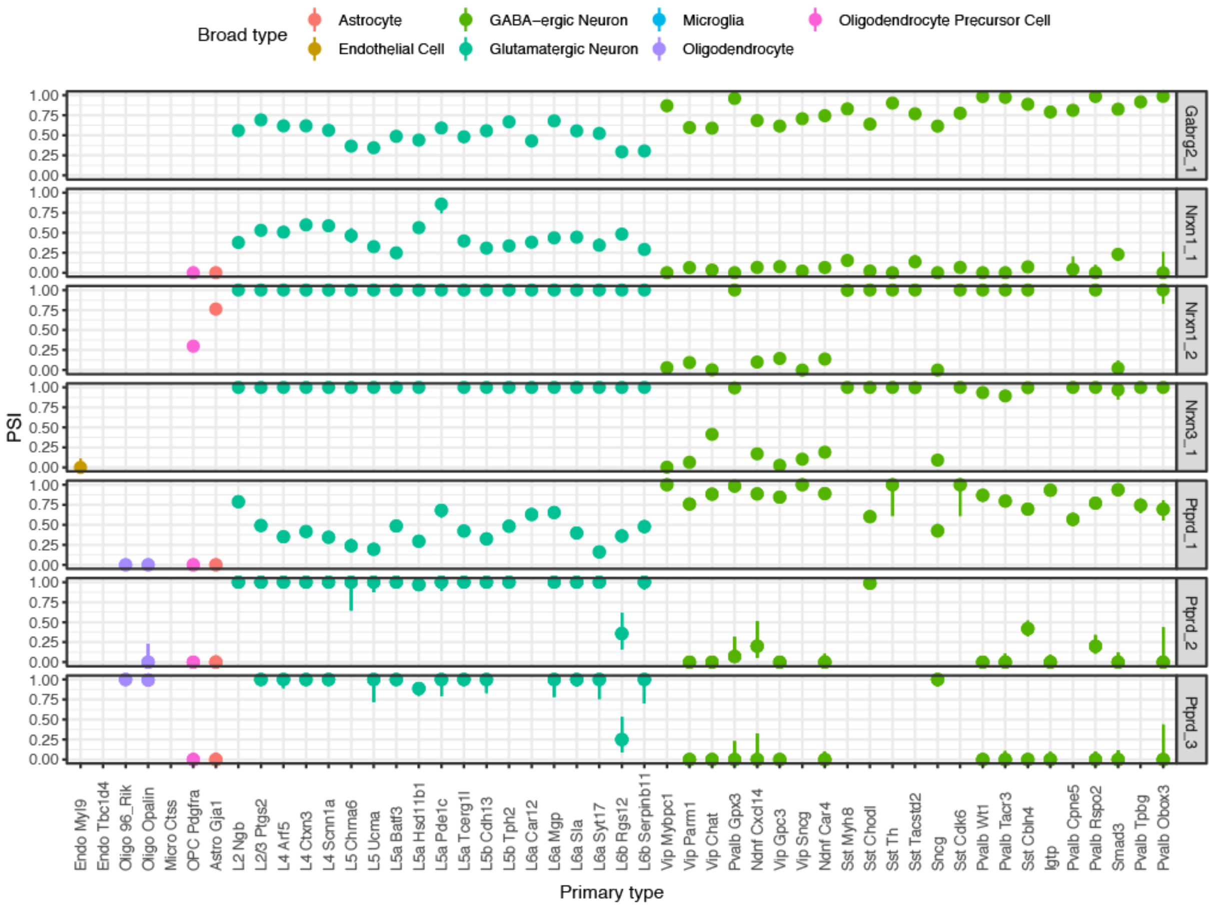

Figure 3.15: Microexon inclusion patterns at synaptic proteins across all core clusters of proteins involved in trans-synaptic interactions. Each panel shows the inclusion pattering of microexons that were found differentially included between GABA-ergic and glutamatergic neurons. Colours indicate the different broad types which each cell-type belong to.

3.3 Methods

3.3.1 Microexon analyses across mouse development using bulk RNA-seq data

As a proof of principle, I applied MicroExonator to 283 RNA-seq datasets, obtained from the ENCODE project (Sloan et al., 2016), corresponding to embryonic and postnatal tissue samples coming from 17 different tissues. For the sequence, I used mm10 mouse genome assembly, obtained from UCSC genome browser database (Karolchik et al., 2003), and as source of annotated splice junctions I used the union of GENCODE Release M16 (Harrow et al., 2006) and VastDB (Tapial et al., 2017). I quantified novel and annotated microexons, Percent of spliced-in (PSI) values, by using MicroExonator's built-in scripts or by using Whippet, which provides a one-step approach for quantify splicing at the splicing node level (Sterne-Weiler et al., 2018). Bi-clustering of samples and microexons were performed applying Ward's minimum variance criterion implemented in R (Müllner and Others, 2013; Murtagh and Legendre, 2014) over a MicroExonator distance matrix where the similarity of the samples was calculated from the PSI values . Moreover, PSI values were also used to perform PPCA using ppca function from pcaMethods R library (Stacklies et al., 2007).

The obtained PPCA loading factors were used to systematically classify microexon clusters. Assuming that PC1 and PC2 are related with variance observed at brain and muscle respectively, loading factors can be used as a proxy to evaluate the tissue-specificity behavior of a given microexon inclusion. Thus, microexons that have high loading factors (>0.03) for PC1 and PC2, were considered as neuromuscular (NM1-3). The ones that only have high loading factors for either PC1 or PC2 were considered as neuronal (N1-4) and muscular (M1-3) respectively. Additionally, one microexon cluster was found with a significant negative loading factor over PC1 (lesser than -0.03), which I considered to be non-neuronal (NN1). I also found microexon clusters that have a consistent inclusion (I1-7) or exclusion

(E1-5) pattern across all the samples and to perform differential microexon inclusion analyses I grouped sample files according to the bi-clustering results.

To perform differential microexon inclusion analyses I grouped sample files according to the bi-clustering results. For each alternative microexon cluster (N1-5, NM1-3 and NN1), baseline and signal sample sets were defined. I quantified splicing nodes using Whippet quantification module (whippet-quant.jl) and I supplied MicroExonator output as input to the Whippet differential inclusion module (whippet-delta.jl). I used both MicroExonator and Whippet quantification to assess changes in microexon inclusion between every signal cluster and its corresponding baseline cluster array. Across the different comparisons, I only considered as significant those microexons which have >0.9 probability of being differentially included and >=0.1 delta PSI values. To further avoid quantification errors, I only selected those microexons that were detected as differentially included using both MicroExonator and Whippet quantification. For each signal cluster I calculated differentially included microexons enrichment using Pearson's chi-squared test with Yates' continuity correction (Yates, 1934). Differentially included microexons were classified accordingly with the tissue composition of signal clusters in which they were found to be differentially included.

I further analyzed the sets of genes that have microexon differentially included in brain, SKM, heart or adrenal gland by building a protein-protein interaction network using STRING (Szklarczyk et al., 2017).

3.3.2 Neuronal mouse dopamine neuron preparation and RT-PCR validations

Mouse embryonic stem cells (mESC) were differentiated into dopamine neurons as previously described (Metzakopian et al., 2015). Briefly, mESCs were first differentiated into Epiblast stem cells (EPI) using fibronectin coated plates and N2B27 basal media (composed of Neurobasal media, DMEM/F12, B27 and N2 supplements, L-glutamine and 2-Mercaptoethanol) supplemented with FGF2 (10mg/ml) and Activin A (25mg/ml). After three passages, EPI were differentiated into dopaminergic neurons using plates collated with poly-L-lysine (0.01%) and

Laminin (10ng/ml) and N2B7 media supplemented with PD0325901 (1mM) for 48hours (Day 0 to Day 2). 3 days later (Day 5), N2B27 media was supplemented with Shh agonist SAG (100nM) and Fgf8 (100ng/ml) for 4 days. Media was then changed to N2B27 media supplemented with BDNF (10ng/ml), GDNF (10ng/ml) and ascorbic acid (200nM) from Day 9 onwards. During neuronal differentiation cells were passaged at Day 3 and Day 9. Cells were collected for qRT-PCR analysis at several stages: mESC, EPI, Day 9 neurons and Day 19 neurons. RNA extraction was performed using the RNeasy Mini Kit (Qiagen) and samples analysed with a QuantStudio 5 PCR system (Thermo Fisher Scientific). These experimental details were provided by Hugo Fernandez, who performed these experiments.

3.3.3 Systematic microexon identification in Zebrafish brain

RNA-seq experiments for the Zebrafish brain tissues across different time points were obtained from (Park et al. 2018) with GEO accession code GSM2971317. Microexon detecting and quantification was performed with MicroExonator using default parameters and taking Ensembl gene predictions 95 and danRer11 genome assembly as an input. To perform a comparative analysis between mouse and zebrafish microexons, I performed a batch coordinate conversion using the liftOver script from USCS utilities (Karolchik et al., 2003), which provides an straightforward, conservative and non-exhaustive way to find conserved microexons.

3.3.4 Single cell analyses

I applied MicroExonator to single cell data from mouse visual cortex (Tasic et al., 2016). In addition to MicroExonator PSI quantification, I also computed PSI using Whippet for all microexons found in this dataset. To compare microexon inclusion rates, I used Whippet to perform an iterative quantification and inclusion analysis. I pooled data coming from 2 neuronal cell-types: GABA-ergic and glutamatergic in pseudo-bulk groups of 5 cells (or fewer for the last group), and repeated this process 10 times. During each iteration, splicing node PSIs values were calculated using whippet-quant.jl. Both single-cell and pseudo-bulk were used to assess differential inclusion of splicing nodes using whippet-delta.jl to obtain average delta

PSI and probability values. Only those that had at least 0.9 mean probability and a mean delta PSI value within single cell delta PSI value +/- 0.25, were considered as significant. Sashimi plots were generated by adapting ggsashimi's code (Garrido-Martín et al., 2018) to display the total number of reads that is supported by each splice site. The read counts were further processed to calculate splice site usage rates.