6 Chapter VI - Discussion and future work

The findings that were described in previous chapters provide some novel insights about two non-canonical splicing features: extremely short exon size (microexons) and the non-canonical DNA and RNA structures associated with splice sites. In the following chapter, these findings are put into perspective and I also highlight future research directions to address yet unsolved questions in this field.

6.1 Development of computational a workflow for reproducible detection and quantification of microexons

The advent of RNA-seq technologies has provided unprecedented opportunities to explore the complexity of vertebrate transcriptomes. Numerous bioinformatics tools can be used to quantify gene expression and many alternative splicing events. However, the detection of splicing events associated with non-canonical features, such as introns with non-canonical dinucleotides, recursive splice sites, back splicing events or microexons, has required the development of specialized bioinformatics methods (Irimia et al., 2014; Li et al., 2015; Parada et al., 2014; Sibley et al., 2015; Wu et al., 2013; Zeng et al., 2017).

In chapter II, I presented MicroExonator, a complete bioinformatic workflow for reproducible discovery and quantification of microexons. Since MicroExonator was implemented using Snakemake as a workflow management system, large volumes of data can be handled using HPC systems while ensuring that the analyses are reproducible. Simulation-based benchmarking results show that MicroExonator has higher sensitivity than widely used RNA-seq alignments tools such as STAR (Dobin et al., 2013) and HISAT2 (Kim et al., 2015) (Fig 2.3, 2.4). Moreover, even though Olego (Wu et al., 2013) has a module specifically dedicated for microexon discovery, sensitivity to find short microexons was comparable to HISAT2, but still inferior to MicroExonator (Fig 2.4c).

While the benchmark results show that MicroExonator has sensitivity improvements particularly for the very short microexons (<10 nt), the evaluation of false-positive microexon rates demonstrates that MicroExonator has significantly higher specificity than START, HISAT2 and Olego at almost the full range size of microexons (Fig 2.4b). These results validate the computational strategies implemented in MicroExonator to reduce the detection of spurious microexons (Fig 2.2), which have been identified as one of the major challenges to perform microexon discovery (Wu and Watanabe, 2005).

6.2 MicroExonator enables large-scale reproducible analyses of microexon splicing

6.2.1 Microexon coordination across neuronal development

Microexon quantitative analysis revealed that the proteins containing microexons form a highly connected PPI network during mouse neuronal development. Moreover, analysis of the topology of the network suggests that the microexons for the most central nodes are included early in development. It is not yet fully understood how this coordination is achieved, but it has been proposed that microexon inclusion relies on an upstream intronic splicing enhancer which is recognized by specific neuronal splicing factors (Gonatopoulos-Pournatzis et al., 2018). However, I also identified a large group of microexons that are constitutively included across murine tissues, suggesting that their inclusion cannot be dependent on tissue-specific factors alone. Instead, our analysis points to a more straightforward explanation as the constitutive microexons have stronger splicing signals than neuronal microexons. Further analysis of neuronal microexon cis-regulatory elements is required to understand how inclusion events are coordinated and why there is a small number of microexons that are progressively excluded through brain development.

The predominant mechanism for regulating alternative splicing events during neuronal development is through RNA binding proteins (Vuong et al., 2016). In the

case of microexons, *SRRM4* and *RBFOX1* have a critical role in coordinating microexon inclusion through brain development, and changes in expression of these splicing factors have been linked to misregulation of alternative splicing events in individuals with autism spectrum disorder (ASD) (Irimia et al., 2014; Li et al., 2015; Voineagu et al., 2011). In fact, alternative splicing changes associated with ASD are enriched in microexons and they are recapitulated in mutant mice haploinsufficient for SRRM4 (Irimia et al., 2014; Quesnel-Vallières et al., 2015). Moreover, a recent genome-wide CRISPR-Cas9 screen has identified two additional factors, SRSF11 and RNPS1, that contribute to SRRM4-dependent microexon regulation, and these genes have also been implicated in ASD and other neurological disorders (Gonatopoulos-Pournatzis et al., 2018). Another example of a protein where imbalances of microexon inclusion have been associated with an elevated risk of ASD is cytoplasmic polyadenylation element-binding protein 4 (CPEB4) (Parras et al., 2018). I found differential inclusion of a *CPEB4* microexon during mouse embryonic brain development, and I also found microexon changes in other protein factors that are involved in mRNA polyadenylation, such as CPEB2, CPEB3 and FIP1L1. All four members of the cytoplasmic polyadenylation element binding (CPEB) family are involved in translational control and have been found to be transcribed in the mouse transcriptome. CPEB transcriptional control has been associated with synaptic plasticity, learning and memory (Turimella et al., 2015). While the role of these microexons in neuronal function and neuropsychiatric diseases remains unexplored, CPEBs function have been associated with ALS and human episodic memory (Downie, 2017; Vogler et al., 2009).

The high degree of conservation of microexons strongly suggests that they are functionally important, however detailed mechanisms of how microexon splicing impacts neuronal function and development have not yet been carried out for most loci. A notable exception is *SRC* where microexon inclusion leads to the production of a well-characterized neuronal splice variant (n-SRC). The *SRC* microexon encodes for a positively charged residue located at an SH3 domain that has been shown to regulate Src kinase activity and specificity (Brugge et al., 1985). From the STRING analysis (Fig 3.3), I found evidence of SRC-dependent phosphorylation of

GIT1, CTNND1 and PTK2 (Chernyavsky et al., 2008; Lim et al., 2002; Wang et al., 2010a). The impact of neuronal microexon alternative splicing for these phosphorylation events remains unknown. However, recent studies show that n-SRC microexon inclusion is required for normal primary neurogenesis and the L1cam dependent neurite elongation (Keenan et al., 2017; Lewis et al., 2017b), implying a strong phenotype. Another central node in the PPI network that is known to undergo microexon alternative splicing changes that are important for axon growth is *L1CAM*, a founding member of L1 protein family. Across the L1 protein family, a sorting signal is included due to 12-nucleotide alternative microexons. In the case of *L1CAM*, the 12-nucleotide microexon mediates clathrin-mediated endocytosis by interacting with adaptor protein complex 2 (AP-2) (Kamiguchi et al., 1998). Our analysis shows that the AP-2 mu subunit (*AP2M1*) is also affected by microexon inclusion through mouse brain development.

6.2.2 Cell-type specific microexon alternative splicing across the mouse visual cortex

Single-cell RNA-seq data are providing unique opportunities to survey cell-specific expression profiles. However, with a few notable exceptions (Arzalluz-Luque and Conesa, 2018; Gokce et al., 2016; Lukacsovich et al., 2019; Zhang et al., 2016), most scRNA-seq analyses have focused on the gene rather than the transcript level. Here, I applied MicroExonator to GABA-ergic and glutamatergic cells from the visual cortex, and to increase power I developed a downstream SnakeMake workflow, snakepool. As many splicing events are undetected in single cell data due to poor coverage, a pooling strategy is necessary to increase the power to identify significant differential inclusion events.

From the analysis with snakepool, 39 microexons were detected as differently included between GABA-ergic and glutamatergic neurons. Fifteen of these cell-type specific microexons are found encoding eleven synaptic proteins. Among these, two alternatively included microexons were found for *PTPRD*, a protein known to have a key role in modulating trans synaptic interactions and having a direct impact on synapse formation (Yamagata et al., 2015a, 2015b). In addition, microexons found in

PTPRD and other proteins involved in transsynaptic protein interactions were found to have distinctive alternative inclusion profiles across GABA-ergic and glutamatergic subtypes (Fig 3.5).

The differential inclusion of microexons could have profound effects on neuronal identity, synapse formation and disease. For example, GABA-ergic neurons were found to have higher inclusion levels of an alternative microexon in GABAa receptor subunit γ (GABRG2) and this alternative splicing event may have significant repercussions for GABA-ergic neuronal function since GABRG2 microexon introduces a phosphorylation site that regulates the GABA activated current (Moss et al., 1992; Ustianenko et al., 2017; Whiting et al., 1990). Misregulation of this alternative splicing event has been associated with schizophrenia in human patients (Huntsman et al., 1998; Ustianenko et al., 2017). However, additional analyses of alternative microexon patterns across neuronal cell-types will be required to fully understand their contribution to neuronal heterogeneity and function.

6.3 Microexon alternative splicing may shape neuronal connectivity

Taken together, the results of the single cell analysis suggest that microexon alternative splicing events may have an influence over synaptic formation across different neuronal subtypes (Fig 3.5). The alternative inclusion of some of these cell-type specific microexons is also shown to be regulated through mouse embryonic development (Fig 3.3). Therefore, it is possible that microexon splicing patterns that are cell-type-specific are established during neuronal development and have a deep impact on the way the neuronal connectome develops to form a mature brain.

Only very recently have bulk RNA-seq analyses started to uncover alternative splicing differences across neuronal subtypes (Furlanis et al., 2019; Saito et al., 2019; Wamsley et al., 2018). These strategies are based on the cell-type specific mRNA isolation based on fluorescence-activated cell sorting or mRNA-pull down approaches, enabling alternative splicing analysis between cellular subtypes by

standard bulk RNA-seq. Wamsley and collaborators used these approaches to study alternative splicing changes that are regulated during cortical interneuron development across specific subcellular types (SST+ and PV+ cINs), where they highlight microexon alternative splicing events in *PTPRD* and *NRXN1* as examples of developmentally regulated alternative splicing events between E18.5 and P4. Moreover, they found that Rbfox1 orchestrates a substantial part of the developmentally regulated alternative splicing events that affect synaptic proteins (Wamsley et al., 2018). Furlanis and collaborators developed a novel approach to perform neuronal cell-type-specific transcriptome profiling, by isolated ribosome-engaged transcripts of genetically defined cortical and hippocampal neuron populations. The analysis of these results showed the existence of an alternative splicing program dedicated to the control of synaptic interactions. Even though they report that only 3.8 - 5.3% of the differentially splicing events across different neuronal cell-types involve microexons, several of the synaptic proteins affected by these events are also affected by differentially included microexon events that I detected in bulk and single-cell RNA-seq data (Fig 6.1)

Figure 6.1: Summary of synaptic genes affected by alternative splicing events across neuronal sub-populations in mouse brain. Cell-type specific alternative splicing events found across brain regions (including forbrain, neocortex and hippocampus) affect spliceosomal proteins. Furlanis and collaborators uncovered alternative splicing programs that might control synaptic interactions and neuronal architecture (Furlanis et al., 2019). * denote genes that are also altered by differentially included microexon events (presented in Chapter III) detected by MicroExonator, either though bulk (red) or single-cell (green) RNA-seq data analysis. Schematics adapted from (Furlanis et al., 2019).

6.4 Non-neuronal microexons

Differential inclusion analyses from RNA-seq samples corresponding to different mouse tissues not only enabled the identification of neuronal microexons, but also microexons that are differentially included in SKM, heart and AG (Fig 2.2f-i). To the best of my knowledge, there are no reports of alternatively included microexons in AG. However, since microexons differentially included in AG highly overlap with microexons differentially included in the brain, they are likely to correspond to neuronal microexons that are included in neuroendocrine cells, known as chromaffin cells, which are derived from the neural crest during embryonic development (Bornstein et al., 2012; Shtukmaster et al., 2013).

Li and collaborators suggest that RBFOX proteins can regulate microexon inclusion across brain, muscle and heart (Li et al., 2015). However, the functional impact of these microexon inclusion events is largely unknown. Several of the alternative microexon events that are common between brain, SKM and heart have an impact

on Mads box transcription enhancer factor 2 (MEF2), particularly on *MEF2A* and *MEF2D* subunits. MEF2 is a transcription factor involved in nervous system development, however RBOX proteins were reported to regulate *MEF2D* alternative splicing events that were required for muscle differentiation (Porter et al., 2018; Runfola et al., 2015). Even though I did not observe any obvious microexon inclusion trend in SKM and heart samples across embryonic development, extensive alternative splicing transitions have been observed during postnatal skeletal muscle development (Brinegar et al., 2017). Thus, microexon developmental changes could be potentially coordinated in later mouse developmental stages that were not included in the RNA-seq experiments that I analysed. Moreover, I found 65 microexons that were differentially included in SKM and/or heart samples (Fig 2.2i), which suggest that additional factors might regulate microexon inclusion in muscular tissues.

6.5 The G-quadruplex formation is enriched in splice sites

Even though B DNA is the most common DNA conformation, different sequence motifs are associated with the formation of non-B DNA structures (Bacolla and Wells, 2004). In Chapter IV, I presented the enrichment analyses of different non-B DNA motifs, of which G4s show the highest enrichment across splice sites. Similar enrichments have been reported by previous in-silico analyses (Maizels and Gray, 2013; Tsai et al., 2014), however I analysed recently published G4-seq data that corroborated the potential of these motifs to form G4 structures near splice sites. More in-depth characterisation of G4 enrichment indicates strong differences between template and non-template strands; while for the upstream exon intronic regions similar level of enrichment was observed for the template and non-template strand, higher non-template G4 enrichment was found for the downstream intronic regions. The high G4 enchments and the strand asymmetries that were observed suggest that G4 positioning around splice sites may be subject to purifying selection, which could be tested by analysing differences on allelic variability of template and non-template G4s that are located in the vicinity of splice sites. Moreover, recent

reports showed that non-template G4 motifs can enhance promoter activity by inducing successive R-loop formation (Lee et al., 2020). This capability of non-template G4 motifs to promote R-loop formation may mediate transcriptional kinetics effects that impact mRNA splicing, however further experiments and analyses will be needed to test this hypothesis.

The evolutionary analyses showed that G4 motifs are a conserved feature in vertebrates, and it may be restricted to mammals and birds (Fig 4.11). However, a more comprehensive evolutionary analysis is needed to fully characterize G4 presence in higher eukaryotic organisms. The presence of additional regulatory mechanisms is in accordance with higher frequencies of alternative splicing events in vertebrates compared to invertebrates (Artamonova and Gelfand, 2007). Moreover, G4s display a higher likelihood of DNA mutations (Du et al., 2014) and as a result they are likely plastic in nature, enabling rapid splicing changes during evolution and the establishment of new functions through alternative splicing and the generation of isoform diversity.

There are certain alternative splicing features that are determinants of exon definition in vertebrates. One of them is splicing strength, which largely influences exon inclusion frequency across isoforms. While strong splice sites are associated with constitutive exons, weaker splice sites lead to suboptimal exon recognition (Luco et al., 2011). Thus, this enables alternative splicing events to be modulated by additional cis-regulatory elements or epigenetic factors (Ast, 2004). Here I show that there is a pronounced enrichment of G4s at weak splice sites and provide evidence for widespread contribution of G4 structures to the regulation of alternative splicing. Xiao and collaborators have also shown splice site strength-dependent association of G-runs across splice sites (Xiao et al., 2009). However, in their analysis they studied independent G-runs, that do not necessarily correspond to G4 motifs, which were found to be more enriched across splice sites with intermediate 5’ splice strength.

6.6 Mechanistic models for G4-dependent modulation

G-runs have been previously reported to be bound by hnRNP F/H, which is known to have a direct influence over splice site recognition (Caputi and Zahler, 2001; Královicová and Vorechovsky, 2006; Marcucci et al., 2007; Mauger et al., 2008; McCullough and Berget, 1997; McNally et al., 2006; Yeo et al., 2004b). Thus, part of the G4 motifs studied in Chapter IV, might overlap with G-runs that are targets for hnRNP F/H binding. However, whether the formation of G4s could enhance or undermine the binding of hnRNP F/H is still a matter of debate. Functional minigene assays suggest mutations that have deleterious effects over G4 formation inhibit exon inclusion by preventing hnRNP F binding, a positive regulator of exon definition (Huang et al., 2017). However, the interpretation of these experiments contradicts previous biophysical evidence which shows that hnRNP F binds preferentially to single-stranded G-tracts, suggesting that G4 formation could have a rather negative effect over exon inclusion (Samatanga et al., 2013). This model of G4 formation as an impediment for hnRNP F binding, is consistent with diverse evidence published by the Burge laboratory which indicates preferential binding of RBPs to unstructured RNA (Dominguez et al., 2018; Lambert et al., 2014; Taliaferro et al., 2016), however more experiments and analysis will be required resolve this conflicting evidence.

Additionally, during transcription G4 formation can be favored by unstranded DNA that is transiently generated inside the transcription fork, favoring DNA G4s at the non-template strand and RNA G4s. In fact, G4 formation has been associated with transcriptionally active promoters, which may lead to genome instability induced by double strand break generation (Hänsel-Hertsch et al., 2016; Marnef et al., 2017). G4 formation can have kinetic effects over RNA PolII by delaying RNA polymerization, in fact gene transcription relies a the co-transcriptional unwinding of G4s by dedicated helicase activity (Chakraborty and Grosse, 2011; Paeschke et al., 2011). Since transcriptional speed has an impact on alternative splicing, G4 formation may lead to alternative splicing regulation through a kinetic control of transcription (Nieto Moreno et al., 2015). Moreover, transcription can induce

RNA/DNA hybrid G-quadruplexes, which can lead to even stronger effects over RNA polymerase progression (Shrestha et al., 2014).

6.7 Depolarisation induced alternative splicing

To investigate the relationship of non-canonical splicing features with dynamic alternative splicing regulation events, we quantified the inclusion of microexons and G4-flanked exons in the context of neuronal depolarization. After KCl-induced depolarization of human and mouse ESC-derived neurons, I observed genome wide changes in alternative splicing patterns, of which around 10% corresponded to exon skipping events. These exon skipping events are much more frequently observed than exon inclusion and they are enriched in microexons and G4-flanked exons (Fig 5.1), suggesting that there is a regulated program of alternative splicing events induced by neuronal depolarization. The prominent role of G4s and microexons suggests that non-canonical splicing features are central to this process. Depolarization neurons leads to a strong increase of intracellular calcium, which is thought to mediate previously reported exon skipping events (Lee et al., 2007; Xie and Black, 2001) (Fig 6.2). However, previous studies have only focused on a handful of genes and to the best of my knowledge this is the first time that thousands of exon skipping events have been shown to be triggered after neuronal depolarization.

The molecular mechanisms by which the increase of intracellular calcium concentration leads to alternative splicing events (particularly exon skipping), are not fully understood. In the case of *NMDAR1* and *KCNMA1* exon skipping events, the increase of intracellular calcium induced by depolarization was reported to activate CaMK IV, a calcium/calmodulin-dependent kinase protein which can regulate splicing selection though specific cis-regulatory elements (CaRREs) which have been shown to be associated with hnRNP L binding sites (Ares, 2007; Lee et al., 2007; Li et al., 2009; Sharma and Lou, 2011; Xie, 2008). However, the increase of intracellular calcium can also lead to CaMK IV-independent alternative splicing changes. For example, exon 19 of RBFOX1 is skipped after depolarization, which leads to an increase of RBFOX1 nuclear localization and induction of

RBFOX1-dependent alternative splicing events (Lee et al., 2009). Another example is NCAM exon 18, which does not respond to the CaMK IV pathway, but instead its skipping is determined by an increase of H3K9ac across exon 18, that is induced after neuronal depolarization. H3K9ac may increase RNA polII elongation rates and has a kinetic effect over splicing of exon 18 of *NCAM*. The presence of G4s near splice sites may modulate the effects of intracellular calcium over alternative exon inclusion events in a CaMK IV-dependent manner. However, G4 structures could also directly respond to changes in intracellular calcium. Miyoshi and collaborators showed that structural transitions from antiparallel to parallel G4 conformations are induced by Ca2+ (Miyoshi et al., 2003). These structural changes may have a direct impact on RNA polII kinetics and splicing, but more experimental evidence will be required to prove or disprove this hypothetical mechanism.

On the other hand, microexons have already been shown to be widely skipped after depolarization of primary cultured hippocampal neurons (Quesnel-Vallières et al., 2016). Voltage-gated calcium channels are key transducers of membrane potential changes of intracellular Ca^{2+} concentration under physiological and experimentally induced conditions (Catterall, 2011). Analysing differential inclusion of microexons, I have found microexons that are differentially included in genes that encode for different subunits of voltage dependent calcium channels (*CACNB1*, *CACNB4*, *CASTSPER2* and *CASTPER2*), and Sodium/calcium exchanger (SLCA1), some of which are annotated members of the “presynaptic depolarization and calcium channel opening” Reactome pathway (Figure 6.2). Thus regulation of microexon inclusion may directly influence intracellular Ca^{2+} concentration changes after neuronal depolarisation and have the potential to regulate part of the alternative splicing events that respond to intracellular Ca^{2+}.

Finally, in several cases, both types of non-canonical splicing features herein studied there were related to the same depolarisation-induced exon skipping events. For example a 30-nt microexon at the UNC13A gene, that was skipped after KCl-induced depolarization, was found to be flanked by a downstream non-template G4 (Fig 5.5). Since UNC13A encodes for a presynaptic protein which plays a key role in glutamatergic transmission, alternative skipping of this exon may have a regulatory

effect over neuronal transmission and neurological diseases (Placek et al., 2019). Another interesting example is the case of a microexon skipping event induced after neuronal depolarization at the *NRXN2* gene, which is flanked by an upstream template G4 (Fig 6.3). This microexon affect an extracellular domain of neurexin-2, implicated in trans-synaptic protein-protein interactions that regulate synaptic formation. Both of these examples are conserved between human and mouse and may be part of a fine-tuned regulatory network of alternative splicing events that coordinate synaptic formation across neuronal populations.

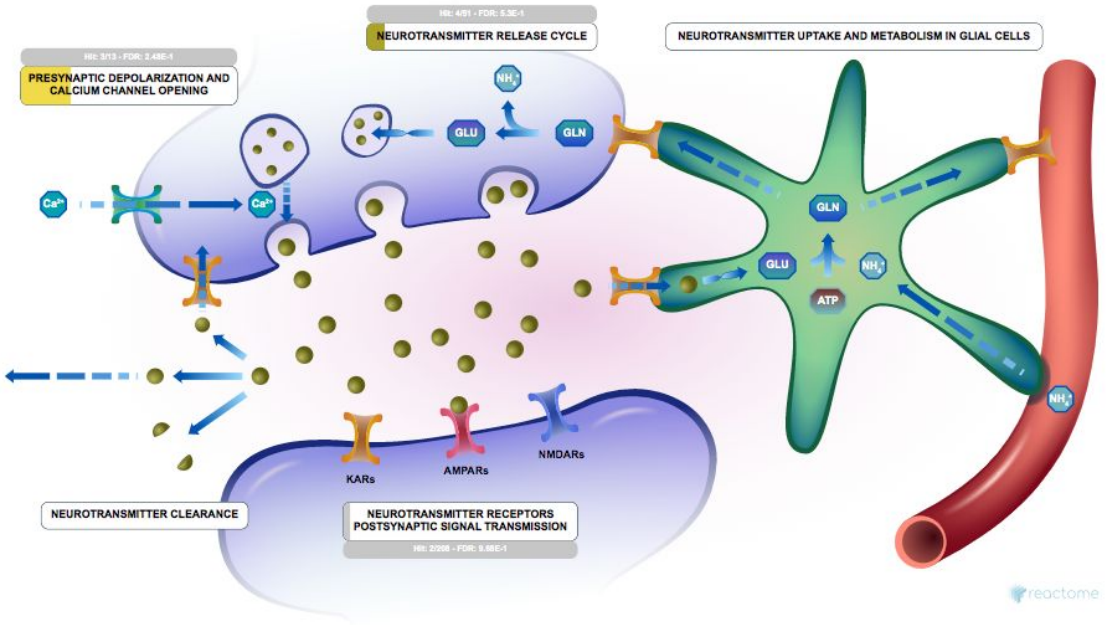

Figure 6.2: Developmentally regulated microexons can have an impact over transmission across chemical synapses. At least two Reactome pathways related with transmission across chemical synapses might be affected microexon inclusion, these includes “Presynaptic depolarization and calcium channel opening” and Neurotransmitter release cycle, where 3/13 and 4/51 genes involved were found to be differentially included through mouse embryonic development using MicroExonator. Bar highlighted in yellow bar indicates statistical enrichment through pathways analysis done using Reactome (Fabregat et al., 2018).

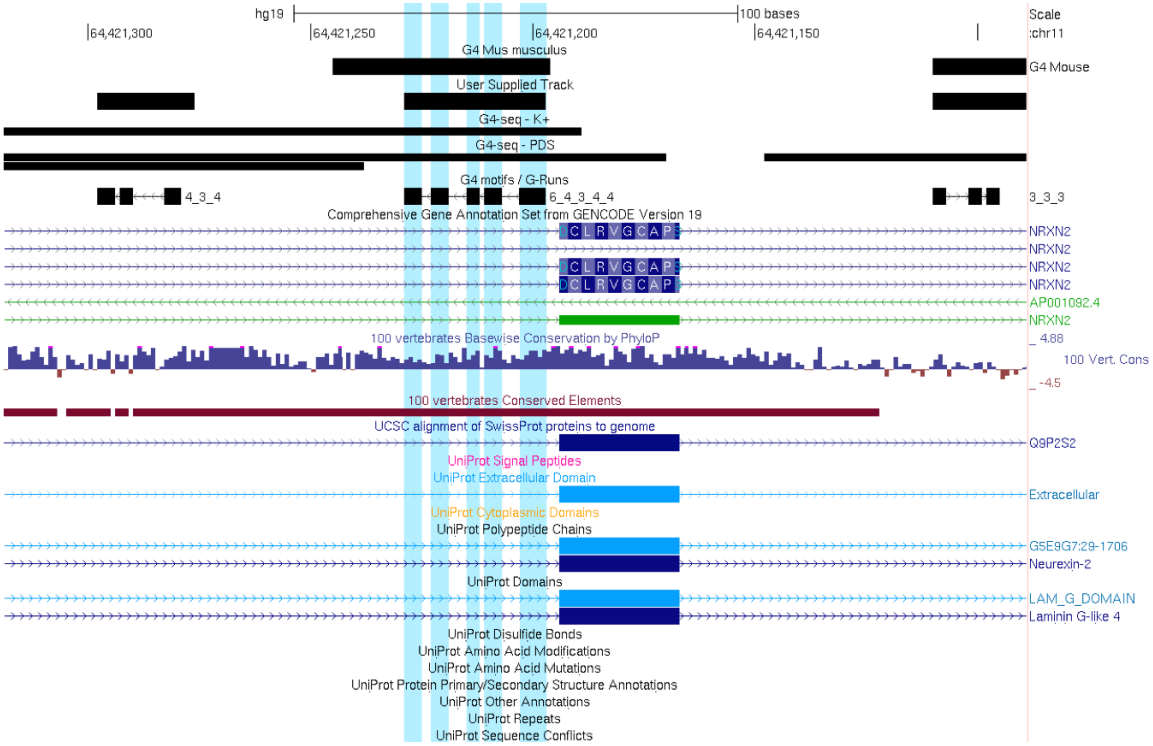

Figure 6.3: Template strand G-quadruplex formation upstream a depolarization-dependent microexon skipping event in neurexin 2. Template G-quadruplex motif is highlighted in blue. G4-seq and additional UCSC tracks are shown.

6.8 Concluding remarks

In this thesis, I have successfully developed novel bioinformatics methods and analysed high throughput sequencing data to characterise microexons and G4-flanked exons. These analyses not only indicate that both short exon size and presence of G4-structures are associated with alternative splicing events, but they also suggest G4s could be potential regulatory features that drive some of the dynamic alternative splicing modulations observed in neuronal cells.

Since microexons are often missing from RNA-seq analyses when standard tools are used, I developed MicroExonator, a novel computational workflow that enables reproducible discovery and quantification of microexons on RNA-seq data. MicroExonator enabled me to perform the integrated analysis of bulk and single-cell RNA-seq data to have an in-depth characterisation of microexons through mouse development and neuronal subtypes. On the other hand, I also explored the possible association of splice sites with sequence motifs that are known to promote the formation of non-canonical DNA and RNA structures. Among the analysed motifs, I found a significant and consistent enrichment of G4 motifs across splice sites of vertebrates.

Further analyses of G4-seq data and experimental validations corroborate *in vitro* formation of G4s near splice sites. Lastly, I analysed the association of microexons and G4-flanked exons to dynamic splicing changes induced by depolarisation stimuli. These analyses showed a transcriptome-wide induction of cassette exon skipping events. Moreover, these changes were enriched in microexons and G4-flanked cassette exons, suggesting the involvement of these non-canonical splicing features in the dynamic regulation of splicing changes upon neuronal depolarisation.

6.9 Future Work

6.9.1 Exploration of large scale RNA-seq sequencing experiments to study alternative splicing of microexons

MicroExonator provides new opportunities to explore large-scale RNA-seq experiments that have been made available to the public domain and this has the potential to enable a comprehensive characterization of microexon alternative splicing patterning across different biological contexts, which may lead to the identification of different cellular pathways associated with alternative splicing of microexons. MicroExonator can be configured to automatically download the data and perform the analysis. Normally, these analyses are limited by the processing power and disk storage that researchers have access to, but the versatile Snakemake workflow management system enables MicroExonator to be compatible with multiple queueing systems (such as LSF or SLURM, implemented at Sanger and Gurdon institute HPC systems respectively). Thus, during the next coming year I will explore several large collections of RNA-seq data to further elucidate the role of the microexon alternative splicing network. In this section will briefly mention the public repositories that I will use for these purposes and what potential insights they may provide.

6.9.1.1 Psychiatric diseases

The network of microexon alternative splicing events have already been shown to be associated with autism. However, a number of microexon splicing regulated events mentioned in this thesis may be involved in the pathogenesis of other neurological diseases. The PsychENCODE project is providing publicly accessible experiments generated from about 1,000 phenotypically characterized disease-affected human post-mortem brains (PsychENCODE Consortium et al., 2015). The large scale multidimensional genomic data generated has already been used to study isoform-level dysregulation associated within autism spectrum disorder, schizophrenia and bipolar disorder, where microexon alternative splicing events have been highlighted and are most likely involved in the pathogenesis of these

disorders (Gandal et al., 2018). Thus, the analysis of these data has the potential to uncover the relevance of microexon alternative splicing under neurological different disease contexts.

1.9.1.2 Primate evolution

The alternative splicing of neuronal microexons is arguably the most conserved network of alternative splicing events so far described (Irimia et al., 2014; Torres-Méndez et al., 2019). Recent research conducted by Torres-Méndez and collaborators has uncovered a novel protein domain, termed as 'enhancer of microexons' (eMIC), which drove the evolution of the neuronal microexon splicing network (Torres-Méndez et al., 2019). While these results trace the evolutionary emergence of this network to bilaterian ancestors, the evolution through primates is currently not characterized. Evolutionary studies of microexons might highlight alternative splicing events that can act as microsurgery of proteins in regulatory regions to coordinate different developmental processes that lead to morphological and functional differences of brains across primates.

In collaboration with Ilias Georgakopoulos-Soares and Professor Nadav Ahituv[9], I am analysing RNA-seq data collected for different non-human primates to uncover the evolutionary trajectory of microexons across primate evolution. We are particularly interested in identifying microexons that are specifically gained or lost in the human lineage and we are going to focus our analysis on data published by The Non-Human Primate Reference Transcriptome Resource (NHPRTR) and genotype-tissue expression (GTEx) project (GTEx Consortium, 2013; Pipes et al., 2013).

6.9.1.2 Functional assessment of RBPs enhanced CLIP and loss-of-function experiments

Recent improvements of CLIP assays used for genome-wide identification of RBP binding sites have led to the development of an enhancer CLIP (eCLIP) protocol (Van Nostrand et al., 2016). These experimental techniques correspond to the

[9] Group leader at Department of Bioengineering and Therapeutic Sciences the University of California San Francisco.

crosslinking and immunoprecipitation of RBP RNA targets and eCLIP experiments that profiled 150 RBPs have enabled the generation of robust splicing regulatory maps (Yee et al., 2019) by the ENCODE consortium (Sloan et al., 2016). Moreover, the ENCODE consortium has recently released a large collection of RBP knockdown followed by RNA-seq experiments across two cell lines. The integration of these resources has been used to explore the role of RBPs across different RNA-processing pathways, including alternative splicing. Furthermore, independent analysis of this data has provided novel insights into the regulation of recursive splicing (Blazquez et al., 2018). Further processing of this data might identify novel regulatory elements across different exon populations, including microexons and G4-flanked exons.

6.9.1.3 Cancer

Alternative splicing defects have been recurrently associated with cancer, however the role of microexon splicing in pathogenic cancer-associated mechanisms is poorly understood. Collin and collaborators have recently identified a microexon alternative splicing event that has functional effects over Cytohesin-1 protein function (Ratcliffe et al., 2019). This corresponds to an evolutionarily conserved 3-nt microexon that induces differential affinity of Cytohesin-1 to triglycine and diglycine, being implicated in selective phosphoinositide recognition and affecting signal transduction pathways related to cancer cell migration. These results showed the potential of microexon splicing research to uncover new mechanisms to drive cancer pathogenesis.

The Catalogue Of Somatic mutations In Cancer (COSMIC) project is currently one of the most ambitious large scale projects at the Sanger Institute (Forbes et al., 2011). COSMIC is the largest and most comprehensive curated catalog of somatic mutations detected in human cancer, providing valuable resources for exploring the impact of somatic mutations in cancer. Recent releases of COSMIC have made available the exome sequencing and RNA-seq data[10] across 1020 cancer cell lines. Therefore, the processing and integration of this data may uncover the effects of

[10] Which for now is only internally available, but it will soon be published.

annotated somatic mutations that are associated with alternative splicing defects of microexons.

6.9.2 Further developments of single cell data analysis methods for microexon splicing analysis

Single cell analyses have been a revolutionary approach to catalog cellular subtypes across different model organisms. Tissues with high cellular heterogeneity, such as brain cortex, have been the target of large scale full-length scRNA experiments released by the Allen Institute for Brain Science (Tasic et al., 2016, 2018). In this thesis I have used part of this data to develop novel approaches to evaluate cell-type specific alternative splicing events (snakemake), particularly focusing on microexon splicing. However, further improvements are required to consolidate this method. Recently developed methods to perform cell-type transcriptomic profiling of neurons using bulk RNA-seq, such as RiboTRAP (Furlanis et al., 2019), could be used to benchmark different single cell approaches to study alternative splicing between two cell-types. This approach is highly attractive since the transcriptome of analogous neuronal populations have been profiled between single cell data generated by Tasic and collaborators and RiboTRAP data generated by Furlanis and collaborators. The integrative analyses of bulk and single-cell RNA-seq experiments might provide significant methodological insights as well as novel cell-type specific microexons and other splicing events.

6.9.3 Study of non-neuronal microexons

The alternative splicing of microexons has been reported to be primarily regulated in neurons, but computational analyses have shown that some of them are also included in heart, SKM and pituitary gland. Moreover quantitative analyses using MicroExonator have also led to the identification of microexons in the adrenal gland (Fig 2.2). I proposed that inclusion of microexons detected in adrenal gland can be due to the presence of chromaffin cells in the adrenal gland, which share the same primordial tissue of origin during embryonic development than other neuronal cells. This hypothesis can now be tested through RNA-seq analysis of isolated chromaffin

cells, which have been recently generated (Chan et al., 2019). Furthermore, the study of other neuroendocrine glands might lead to the identification of microexon inclusion in other non-neuronal tissues, such as different components of the gastrointestinal tract, gallbladder, pancreas and thyroid.

Even though microexon inclusion has been reported across SKM and heart their functions are largely unknown. Moreover, other types of muscle, such as smooth muscle have not yet been studied. Collaborations with Professor Christopher WJ Smith will be initiated to explore microexon inclusion in smooth muscle and determined if RBPMS, a newly identified splicing factor that control smooth muscle splicing events, is implicated in microexon alternative splicing (Nakagaki-Silva et al., 2019).

6.9.4 Tissue-specific splicing of G4-flanked exons

The results that I have presented in Chapter V indicate that microexon and G4-flanked exons are enriched in alternatively included exons induced by depolarisation stimuli. In the case of microexons, alternative splicing events are strongly associated with neuronal alternative splicing programmes. However, in the case of G4-flanked exons we have not yet systematically explored their inclusion across different tissues and cell-types. Thus, similar analyses to the ones I have conducted for microexons are required to determine if G4-exons are in general associated with alternative splicing, or if they are particularly associated with tissue-specific or cell type-specific splicing events.

6.9.5 Elucidating mechanisms of G4-mediated modulation of alternative splicing

The formation of G4 structures can affect alternative splicing by altering the binding potential of RBPs to their target sites. As mentioned in section 1.6.2.1, there is still controversy about whether G4 formation would promote or block the binding of RBPs to the mRNA. Since there is an increasing amount of crosslinking and immunoprecipitation (CLIP) experiments that is being generated to determine the binding sites of multiple RBPs (Louie et al., 2018; Yee et al., 2019), I plan to do a

systematic analysis of RBP binding sites across splice sites and G-quadruplexes. This analysis will provide an unbiased approach to find novel G4 interactors that may play a functional role over alternative splicing regulation.

Another non-exclusive possibility is that G4 formation leads to kinetics effects that regulate alternative splicing. G4s have already been reported to have an impact in RNA polymerase speed and kinetics, but their effects on splicing modulation remain undescribed. However, polymerase speed is already known to regulate splicing and control the recognition of weak splice sites (Luco et al., 2011). Thus the integration of genome-wide transcription pausing with G4-seq and RNA-seq data might provide new mechanistic insight about the kinetic effects of G4 formation in alternative splicing.

6.9.6 Machine learning for motif discovery

Machine learning approaches have been implemented to uncover novel *cis*-regulatory elements that control tissue-specific alternative splicing (Barash et al., 2010; Zhang et al., 2019b). Recent doctoral work of Nicholas Lee[11] has led to the development of a novel convolutional neural network approach to find regulatory motifs across different quantitative transcriptome experiments. Collaborative work between Nicholas Lee, Jacob Hepkema[12] and I have led the identification of novel microexon regulatory motifs that can quantitatively predict the inclusion patterns observed across mouse brain development. Thus, further inspection of these results can lead to the discovery of novel cis-regulatory elements involved in microexon alternative splicing.

[11] PhD candidate at the Sanger Institute and University of Cambridge, who has also been under the supervision of Martin Hemberg.
[12] Master student from Utrecht University, who have been doing analyses at Hemberg's laboratory as an internship work.