Computational analyses of non-canonical architectural and structural features associated with alternative splicing

UNIVERSITY OF CAMBRIDGE

Submitted for the degree of Doctor of Philosophy

by

Guillermo Eduardo Parada González

University of Cambridge

Wellcome Sanger Institute

Homerton College

September 2019

Preface

The dissertation is submitted for the degree of Doctor of Philosophy.

I declare this is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated in the text. This document, in whole or in parts, has not been submitted for any other degree or diploma.

It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Guillermo Eduardo Parada González

Cambridge, UK.

July 2020

Computational analyses of non-canonical architectural and structural features associated with alternative splicing

Guillermo Eduardo Parada González

Summary

Splicing of nuclear introns is catalysed by the spliceosome, one of the most complex macromolecular machines currently known. Even though the canonical splicing signals that drive the precise recognition of splice sites are well-characterised, recent advances in transcriptome profiling technologies and computational method development have enabled widespread identification of non-canonical splicing features. Non-canonical splicing is highly associated with dynamic splicing regulation, and occurs most prevalently in neuronal tissues. In this present work, I have investigated two types of non-canonical features that are related to atypical exon-intron structures and DNA/RNA conformations.

First, I studied a group of extremely small exons, known as microexons (≤30 nucleotides), which were shown to be part of an evolutionarily conserved network of neuronal alternative splicing events that play essential roles in neuronal development. Since standard RNA-seq tools cannot efficiently detect microexon splice sites, I developed MicroExonator, a novel pipeline for reproducible de novo discovery and quantification of microexons. As a proof of principle, I analysed microexon alternative inclusion patterns across 289 RNA-seq samples coming from eighteen different tissues across a wide range of mouse embryonic and adult stages. I detected 2,938 microexons, 343 of which are differentially spliced throughout mouse embryonic development, including 35 that are not present in mouse transcript annotation databases. Unsupervised clustering of microexons alone segregates brain tissues by developmental time and further analysis suggest a key function for microexon inclusion in axon growth and synapse formation. Moreover, I developed a module to adapt MicroExonator splicing analysis to single-cell RNA-seq samples that I used to analyse data from the mouse visual cortex. As a result, I found 39 microexons that are differentially included between glutamatergic and gabaergic neurons, fifteen of which are found in genes that encode synaptic proteins.

The second type of non-canonical features that I studied are sequences associated with non-B DNA structures and possibly atypical RNA conformations. I analysed the enrichment of different non-B DNA motifs across splice site sequences. The strongest and most consistent enrichments were found for G-quadruplex motifs, which are enriched ~3-fold both upstream and downstream of splice junctions. Further analysis of G4-seq experiments corroborated the enriched motifs detected at splice sites leads to *in-vitro* G-quadruplex formation. Moreover, enrichment analyses of G-quadruplex motifs and G4-seq experiments across multiple species suggest that the association of G-quadruples to splice sites is a property restricted to mammals and birds. Interestingly, I found stronger enrichment of G-quadruplexes associated with weak splice sites, suggesting that they could function as cis-regulatory elements of alternative splicing events.

Finally, to explore if microexons and exons flanked by intronic G-quadruplexes were involved in dynamic splicing changes, I analyse alternative splicing events induced by depolarisation treatments in human and mouse neurons. I found a widespread cassette exon skipping response after neuronal depolarization, which was particularly enriched in microexons and exons flanked by G-quadruplexes motifs. Taken together, these results suggest that non-canonical splicing features are an important regulatory mechanism of alternative splicing. Further characterisation of non-canonical splicing might provide a better understanding of fine-tuned alternative splicing mechanisms, in particular in the context of neuronal development and heterogeneity.

This thesis is dedicated to my parents, both Mabel and Guillermo, who have devoted a big part of their life to raise me as the man who I am today. They not only provided me unconditional love and patience, but also the opportunities that enabled me to find my passion in life. I would like to thank my grandmother Mercedes for sparking my scientific curiosity at a very young age by buying my fun scientific books when I was a child. Finally to my grandfather Guillermo, now resting in peace, the first academic of our family, with whom I had very passionate discussions about science and life that I will never forget.

Acknowledgements

I want to thank Dr Martin Hemberg and Prof Eric Miska for supervising me during my PhD. I much appreciate the support they have provided me during this period, which enabled me to take full advantage of my privileged position as a PhD student of this prestigious institution. Members from both Hemberg and Miska lab contributed significantly towards my initial learning and development. Particularly, I would like to thank Dr Tallulah Andrews, Dr Vladimir Kiselev and Dr Tomás Di Domenico. I also want to sincerely thank Dr Ilias Georgakopoulos-Soares, the first graduated PhD from Hemberg lab, for all of his support, collaboration and friendship. I also want to acknowledge Dr Sarah Teichmann and prof Chris Smith for their critical feedback of my doctoral research, as part of my thesis committee, and also to Prof Chris Ponting and Dr Jen Harrow for accepting to read and evaluate this thesis.

I am really grateful to my parents, Mabel and Guillermo, for their unconditional love and support, and to Isabel, my beloved girlfriend, for her much needed support during the toughest moments of PhD. I also thank all other family members and friends from Chile, particularly the CDP group, Joaquin, Jacqueline and Maria Jose, who were always there for me. I would also like to thank my dear friends from Cambridge, particularly Dr Jenkinks, Dr Fryer, Dr Singh, as well as future doctors Eijsbouts, De Jonghe and Kosałka. My experience as a PhD student would not have been the same without them.

Finally, I would like to acknowledge other people who played an important role to motivate me in my early years to pursue a career as a researcher. I thank Dr Eduardo Ravanal, who was my high school biology teacher, for teaching me the basic concepts of molecular biology and pushing me towards academia. I also thank Prof Katia Gysling for all her support during my undergraduate training, which enabled me to successfully get a position as a PhD student of this university. Finally, I would like to thank Dr Roberto Munita for the training he provided during my undergraduate research, which sparked my interest in computational biology.

List of figures

Figure 1.1: Splicing mechanism. ……6
Figure 1.2: Spliceosomal core signals and assembly. ……9
Figure 1.3: Alternative splicing events types..……..14
Figure 1.4: G-quartet and G-quadruplex structure..……...32
Figure 1.5: Neuronal splicing code..……..39
Figure 1.6: Development of novel bioinformatic methods have enabled the detection of different non-canonical splicing..…….……44
Figure 2.1: Overview of the MicroExonator workflow..…..……..55
Figure 2.2: Quantitative microexon exon filtering...…….56
Figure 2.3: Ground truth generation for the assessment of microexon exon discovery modules ……...57
Figure 2.4: Evaluation of microexon discovery performance of RNA-seq aligners and MicroExonator using synthetic data…..……59
Figure 3.1: Microexon inclusion through mouse embryonic development. ……....67
Figure 3.2: Microexon PSI biclustering ..……..69
Figure 3.3: PPCA loading factors across microexon clusters…..……71
Figure 3.4: Inclusion properties of microexon clusters……...….73
Figure 3.5: Microexon PSI values across all identified microexon clusters……....75
Figure 3.6: Mean PSI values across neuronal and neuromuscular microexons……....76
Figure 3.7: An overview of Whippet's computational workflow to quantify alternative splicing events……...78
Figure 3.8: Differential inclusion analysis performed MicroExonator and Whippet quantification outputs show similar trends……..80
Figure 3.9: Differential inclusion analysis of microexons ……...82
Figure 3.10: Differences in PSI score between adrenal gland, brain MHN and forebrain tissues. ……....83
Figure 3.11: Microexon protein-protein interaction network. ……..…...87
Figure 3.12: Discovery of novel microexons in mouse and zebrafish. ……89
Figure 3.13: Differences between unpooled and pooled methodologies to assess microexon splicing changes in single cell data. ……...92
Figure 3.14. Differential alternative splicing analysis of microexons between glutamatergic and GABA-ergic neurons. ……....94
Figure 3.15. Microexon inclusion patterns at synaptic proteins across all core clusters of proteins involved in trans-synaptic interactions. ……...95
Figure 4.1: Landscape of non-B DNA motifs across human splice sites. ……103
Figure 4.2: Non-B DNA motif enchiment varies with splice strength. ……....104
Figure 4.3: Analysis of high-resolution G4-seq data validates in-silico enrichment of G4 motifs…..……..……107
Figure 4.4: Splice site strength and distribution of G4 motifs at splicing sites ……....109
Figure 4.5: Characterisation of G4 motifs across splicing junctions. ……..111
Figure 4.6: G4 enrichment patterns are consistent across different G4-seq experiments….
……113

Figure 4.7: Template and nontemplate splice site G4 enrichment across gene body. ……115
Figure 4.8: Enrichment of G4 of different G-run lengths. ……116
Figure 4.9: G4s are enriched in short intron splice sites. ……118
Figure 4.10: G4 enrichment at short intron splice sites is driven by GC-content ……120
Figure 4.11: G4 motifs are enriched in a subset of vertebrates. ……122
Figure 4.12: Cross specie G4-seq analyses validate findings *in-silico*……...124
Figure 5.1: Depolarization induces genome-wide exon skipping of cassette exons.……...134
Figure 5.2: Mouse cortical neuronal depolarization experiments shows G4-mediated alternative splicing patterns consistent with human results..……136
Figure 5.3: Cassette exons exhibit a strong exon exclusion pattern after KCl-induced depolarization.....……...138
Figure 5.4: Non-template G-quadruplex motif downstream an alternatively included SLC6A17 exon. ……139
Figure 5.5: Non-template G-quadruplex motif downstream an alternatively included UNC13A microexon……140
Figure 5.6: Non-template G-quadruplex motif downstream an alternatively included NAV2 exon.……...141
Figure 6.1: Summary of synaptic genes affected by alternative splicing events across neuronal sub-populations in mouse brain. ……..151
Figure 6.2: Developmentally regulated microexons can have an impact over transmission across chemical synapses. ……....158
Figure 6.3: Template strand G-quadruplex formation upstream a depolarization-dependent microexon skipping event in neurexin 2. ……159

List of abbreviations

AG	Adrenal gland
CaMK	Calmodulin-dependent protein kinase
CaRRE	CAMK IV-responsive RNA element
cDNA	Complementary DNA
CE	Core exon
CRISPR	Clustered regularly interspaced short palindromic repeats
CSG	Contiguous splice graph
DNA	Deoxyribonucleic acid
DPC	Days post conception
EEEJ	Exon-microexon-exon junctions
EJC	Exon junction complex
EPI	Epiblast stem cell
EST	Expressed sequence tag
G4	G-quadruplex
GO	Gene ontology
GTF	Gene transfer format
hnRNP	Heterogeneous nuclear ribonucleoprotein
mESC	Mouse embryonic stem cell
mRNA	Messenger RNA
NMD	Nonsense-mediated decay
NMDA	N-Methyl-D-aspartic acid
PCR	Polymerase chain reaction
PDS	Pyridostatin
PPCA	Probabilistic principal component analysis
PPI	Protein-protein interaction network
PSI	Percent spliced in
PTC	Premature stop codon
RBP	RNA-binding protein
RNA	Ribonucleic acid
RT	Reverse transcriptase
RUST	Regulated unproductive splicing and translation
SKM	Skeletal muscle
snRNP	Small nuclear ribonucleoprotein
ss	Splice site
UTR	Untranslated region

Index

1 Chapter: Introduction **4**
1.1 Splicing; a pivotal step of eukaryotic RNA-processing 6
1.1.1 Spliceosomal machineries 7
1.1.2 Canonical nuclear eukaryotic splicing 8
1.1.3 Core spliceosomal splicing signals 8
1.1.4 Spliceosome assembly and catalysis. 10
1.1.5 Precise recognition of splice sites 11
1.1.5.1 Cis-acting regulatory elements 12
1.1.5.2 Intron and exon definition 12
1.2 Widespread alternative splicing expands transcriptome and proteome diversity in vertebrates 13
1.2.1 Unproductive splicing events 15
1.2.2 Global assessment of alternative splicing and its impact over the proteome diversity 17
1.2.2.1 Mass-spectrometry based assays: Futile alternative splicing events or lack of sensitivity? 17
1.2.2.2 Alternative splicing events rewire protein interaction networks across tissues 18
1.3 Fine-tuned control of alternative splicing 19
1.3.1 Features associated with alternative splicing events 19
1.3.1.1 Splice site strength 20
1.3.1.2 Gene-architecture effect on alternative splicing 20
1.3.1.3 Epigenetic modulation 21
1.3.1.4 Effect of secondary structures 22
1.4 Non-canonical splicing feature effects over alternative splicing 23
1.4.1 Unusual splice sites 24
1.4.1.1 A minority group of introns is processed by a dedicated parallel spliceosomal machinery 24
1.4.1.2 Non-canonical splice sites 24
1.4.1.2.1 XBP1 intron is the only known nuclear intron that is not processed by the spliceosome 25
1.4.1.3 Cryptic-splice sites 25
1.4.1.4 U2AF65 independent splicing 26
1.4.2 Non-canonical intron-exon structures 26
1.4.2.1 Analysis of short and ultra-short introns 27
1.4.2.2 Microexons 28
1.4.2.3 Recursive splicing 29
1.5 Non-canonical nucleic acid structures 29
1.5.1 G-quadruplex formation 30
1.5.2 R-loop formation 33
1.6 Deciphering the non-canonical splicing code and its implications in tissue-specific splicing 34
1.6.1 Transcriptomic revolution 34
1.6.2 Alternative splicing tissue-specific code 35
1.6.2.1 Canonical and non-canonical neuronal splicing code 35
1.6.2.1.1 Sequence motif code 35
1.6.2.1.2 Architectural code 36
1.6.2.1.3 The RNA structural code 37
1.6.3 Non-canonical splicing detection and quantification using RNA-seq data 40
1.6.3.1 Identification of neuronal non-canonical splicing events using RNA-seq data 41
1.6.3.1.1 Recursive splicing detection 42
1.6.3.1.2 Identification of circRNAs 42
1.6.3.1.3 Discovery and quantification of microexons 43

1.6.3.2 Genome-wide evaluation of non-canonical RNA-structures effects over alternative splicing 45
1.6.3.2.1 Genome-wide detection of G4 formation and its impact over alternative splicing modulation 45
1.7 Research aims 47

2 Chapter II: Reproducible RNA-seq processing for detection and quantification of microexons 48
2.1 Introduction 48
2.1.1 Computational methods for discovery and quantification of microexons using RNA-seq data 49
2.1.2 Reproducible bioinformatics analysis using workflow manager platforms 51
2.1.3 Computational environments 52
2.2 Results 53
2.2.1 Development of a reproducible bioinformatic workflow to discover and quantify microexons in RNA-seq data 53
2.2.2 Benchmarking of computational methods for microexon discovery 57
2.3 Methods 60
2.3.1 Annotation guided microexon discovery using RNA-seq data 60
2.3.2 Quantification of microexon inclusion 61
2.3.3 Filtering of spurious intronic matches 61
2.3.4 RNA-seq simulation 62

3 Chapter III: Microexon quantitative analyses across mouse brain development and visual cortex 64
3.1 Introduction 64
3.2 Results 66
3.2.1 Microexon inclusion changes dramatically over mouse embryonic development 66
3.2.2 Microexon alternative splicing is coordinated throughout embryonic development 84
3.2.3 MicroExonator enables the identification of novel neuronal microexons 87
3.2.4 Identification of microexons in zebrafish brain. 88
3.2.5 Cell type specific microexon inclusion in mouse visual cortex. 90
3.3 Methods 96
3.3.1 Microexon analyses across mouse development using bulk RNA-seq data 96
3.3.2 Neuronal mouse dopamine neuron preparation and RT-PCR validations 97
3.3.3 Systematic microexon identification in Zebrafish brain 98
3.3.4 Single cell analyses 98

4 Chapter IV: Analysis of non-B DNA motifs across splice sites 100
4.1 Introduction 100
4.2 Results 101
4.2.1 Genome wide analysis of non-B DNA motifs across splice sites 101
4.2.2 G4 enrichment analyses of splice sites 105
4.2.2.1 G4s are enriched at weak splice sites 108
4.2.2.2 G4s are preferentially found on the non-template strand 110
4.2.2.2.1 Replicated effects are found in independent G4-seq experiments 112
4.2.2.2.2 Template and non-template G4 enrichment patterns across gene body 114
4.2.3 Gene architectural features associated with G4-exons 117
4.2.3.1 G4s are enriched for short introns 117
4.2.3.1.1 GC-content controlled associations of G4 motifs and intron size 119
4.2.3.2 G4 are not enriched in microexons 121
4.2.4 Abundance of G4s at splice sites has emerged during vertebrate evolution 121
4.3 Materials and Methods. 125
4.3.1 Genome and gene annotations processing 125
4.3.2 Genomic datasets. 125
4.3.2.1 Non-B DNA motifs. 125

4.3.3 G4-seq data 127
4.3.4 Relationship between G4s and exon / intron length 127
4.3.4.1 G4s and relationship to exon number 128
4.3.4.2 Relationship between G4s, splicing strength score and intron length 128

5 Chapter V: Dynamic non-canonical splicing responses to neuronal depolarization stimuli 129
5.1 Introduction 129
5.2 Results 131
5.2.1 Dynamic splicing responses to neuronal depolarization are associated with non-canonical features 131
5.2.1.1 Depolarization triggers genome-wide cassette exon exclusion events that are highly enriched in microexons 131
5.2.1.2 Dynamic splicing responses to neuronal depolarization are associated with G4s proximal to splice sites 132
5.2.2 Case study of G4 associated with depolarization induced exon skipping events that are evolutionarily conserved 137
5.2.2.1 Candidate selection 137
5.2.2.2 G4 motif sequences found at SLC6A17, UNC13A and NAV2 promote the formation of G4 structures in vitro 142
5.3 Materials and Methods. 143
5.3.1 Comparative analysis of RNA-seq experiment. Differential exon inclusion following depolarisation. 143

6 Chapter VI - Discussion and future work 145
6.1 Development of computational a workflow for reproducible detection and quantification of microexons 145
6.2 MicroExonator enables large-scale reproducible analyses of microexon splicing 146
6.2.1 Microexon coordination across neuronal development 146
6.2.2 Cell-type specific microexon alternative splicing across the mouse visual cortex 148
6.3 Microexon alternative splicing may shape neuronal connectivity 149
6.4 Non-neuronal microexons 151
6.5 The G-quadruplex formation is enriched in splice sites 152
6.6 Mechanistic models for G4-dependent modulation 154
6.7 Depolarisation induced alternative splicing 155
6.8 Concluding remarks 160
6.9 Future Work 161
6.9.1 Exploration of large scale RNA-seq sequencing experiments to study alternative splicing of microexons 161
6.9.1.1 Psychiatric diseases 161
6.9.1.2 Functional assessment of RBPs enhanced CLIP and loss-of-function experiments 162
6.9.1.3 Cancer 163
6.9.2 Further developments of single cell data analysis methods for microexon splicing analysis 164
6.9.3 Study of non-neuronal microexons 164
6.9.4 Tissue-specific splicing of G4-flanked exons 165
6.9.5 Elucidating mechanisms of G4-mediated modulation of alternative splicing 165
6.9.6 Machine learning for motif discovery 166

7 References 167