Contents

Glossary **xxv**

1 Introduction **1**

1.1 Alternative Splicing 3

1.2 High-throughput technologies to characterise transcriptome complexity 7

1.2.0.1 PacBio Iso-Seq sequencing 9

1.3 Approaches for isoform quantification and differential isoform usage 14

1.4 Functional impact of isoform regulation 16

1.4.0.1 Functional impact on protein properties 16

1.4.0.2 Functional impact on UTR properties 18

1.4.0.3 Nonsense-mediated decay 19

2 Motivation, Aims, and Contributions **21**

2.1 Motivation 23

2.2 Aims 25

2.3 Main contributions 27

2.3.1 Journal papers 27

2.3.2 Conferences 28

2.3.3 Awards 30

2.3.4 Software 31

2.3.5 Master's Thesis Supervisions 31

2.3.6 Teaching 31

3 Extensive characterization and quality control of long-read sequencing transcriptomes 33
3.1 Introduction 35
3.2 Data 38
3.2.1 Neural System in mouse 38
3.2.2 Public datasets 39
3.2.2.1 Maize PacBio 39
3.2.2.2 MCF-7 Human PacBio 39
3.2.2.3 B-cell mouse nanopore 40
3.3 Methods 41
3.3.1 Transcriptome definition using Iso-Seq PacBio long-reads 41
3.3.2 Iso-Seq PacBio evaluation of isoform quantification and detection 42
3.3.3 Classification of transcripts to describe long-read captured novelty 43
3.3.4 Extensive isoform characterisation as a means for quality control 45
3.3.5 Using quality control features to build a filter of isoform artefacts 47
3.3.6 Open reading frame prediction benchmarking and assessment of UTR/ORF variability in PacBio-defined transcriptomes 49
3.4 Results 51
3.4.1 PacBio Iso-Seq sequencing quality 51
3.4.2 Transcriptome complexity and transcript full-lengthness assessment across alternative pipelines 51
3.4.3 Characterisation of ToFU-defined novel calls reveals enrichment in artefacts 61
3.4.4 Experimental validation of ToFU results verifies the presence of novel-isoform artifacts 65

3.4.5 Machine learning enables accurate filtering of novel-isoform artefacts . . . 67

3.4.6 PacBio sequencing unable to accurately quantify low-medium expressed isoforms but capturing most transcriptional signal 71

3.4.7 Novel transcripts have a major impact on the accuracy of transcriptome quantification by short reads . . . 73

3.4.8 Open reading frame prediction in long-read defined transcriptomes . . . 76

3.4.9 Open reading frame diversity generated by novel long-read defined isoforms . . . 77

3.4.10 SQANTI tool . . . 79

3.4.10.1 SQANTI quality control . . . 80

3.4.10.2 SQANTI filter . . . 81

3.4.10.3 Diagnostic plots . . . 81

3.4.11 Generalization of the SQANTI approach . . . 82

3.5 Discussion . . . 85

4 In-silico annotation of isoforms with functional and regulatory features 89

4.1 Introduction . . . 91

4.2 IsoAnnot pipeline . . . 94

4.2.1 Input data . . . 94

4.2.2 Functional annotation at transcript isoform resolution . . . 94

4.2.2.1 Cis-acting UTR regulatory elements . . . 96

4.2.2.2 Upstream open reading frame prediction . . . 96

4.2.2.3 Repeats and low-complexity elements . . . 97

4.2.2.4 miRNA binding sites . . . 97

4.2.2.5 RNA binding protein binding sites . . . 101

4.2.2.6 Polyadenylation signals . . . 102

4.2.2.7 Nonsense-mediated decay . . . 103

4.2.3 Functional annotation at protein isoform resolution . . . 103

4.2.3.1 Pfam domains . . . 103

4.2.3.2 Transmembrane domains 104
4.2.3.3 Signal peptide 104
4.2.3.4 Coiled regions 104
4.2.3.5 Disordered regions 105
4.2.3.6 Nuclear localization signals 105
4.2.3.7 Coordinate-based and in-frame transference of protein functional features 106
4.2.4 Annotation of non-positional functional information 111
4.2.5 IsoAnnot output 111
4.3 Results 112
4.3.1 Functional annotation of PacBio-defined neural transcriptomes 112
4.3.2 Functional annotation of reference transcriptomes and proteomes 121
4.4 Discussion 125

5 Comprehensive framework for the functional analysis of alternative isoform usage 129
5.1 Introduction 131
5.2 Methods 134
5.2.1 Input data 134
5.2.2 Module 1: Isoform functional diversity 134
5.2.2.1 Structural diversity 135
5.2.2.2 Feature diversity 137
5.2.2.3 Overall rate of diversity 138
5.2.3 Module 2: Transcriptome dynamics 139
5.2.3.1 Differential expression 139
5.2.3.2 Differential isoform usage 140
5.2.4 Module 3: Functional impact triggered by isoform regulation 145
5.2.4.1 Differential feature inclusion 145
5.2.4.2 Differential polyadenylation and UTR lengthening 148
5.2.5 TappAS software implementation 152

5.3 Data 154
5.4 Results 155
5.4.1 The impact of neural trancriptome complexity on functional diversity 155
5.4.2 Multi-layer analysis of alternative isoform usage 164
5.4.3 Impact of differential isoform usage on functional properties169
5.4.4 Impact of APA events in UTR modulation and containing features 177
5.5 Discussion 181

6 Functional consequences of differential isoform usage in neural fate determination 185

6.1 Introduction 187
6.2 Methods 189
6.2.1 Experimental design 189
6.2.2 RNA-seq by single molecule and short-read sequencing . 189
6.2.3 De novo discovery of neural isoforms by Iso-Seq and transcriptome curation 190
6.2.4 Isoform quantification and normalisation 191
6.2.5 Principal component analysis for lineage characterisation 191
6.2.6 Relevance of novel isoforms in defining cell identity 192
6.2.7 Transcriptional and post-transcriptional dynamics in neural differentiation systems 193
6.2.8 Isoform functional diversity in neural differentiation systems 194
6.2.9 Differential feature inclusion between glial and neuronal differentiation 195
6.2.10 Differential polyadenylation in differentiating glial and neuronal cells 195
6.3 Results 197
6.3.1 Widespread novel post-transcriptional diversity in neural systems captured by long-read sequencing 197

6.3.2 System characterisation defines oligodendrocyte and motor neuron progenitors as the most mature differentiation stages 201
6.3.3 Novel isoforms are not prevalent but define cell-lineage and differentiation dynamics 203
6.3.4 Membrane trafficking among the strongest processes specifically regulated by alternative isoform usage in neural determination 207
6.3.5 Functional impact of differential isoform usage on neural fate determination 214
6.4 Discussion 231

7 Conclusions 239

Appendix 1: SQANTI attributes at transcript level 247

Appendix 2: SQANTI attributes at splice junction level 251

References 255