INTEGRATIVE COMPUTATIONAL APPROACHES TO STUDY PROTEIN-NUCLEIC ACID INTERACTIONS

ANOB MAULI CHAKRABARTI

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
of
UNIVERSITY COLLEGE LONDON.

Department of Genetics, Evolution and Environment
UCL Genetics Institute
University College London

February 2020
I, Anob Mauli Chakrabarti, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.
ABSTRACT

Interactions between proteins and nucleic acid molecules are central to the cellular regulation and homeostasis. To study them, I employ a wide range of computational analysis methods to integrate genomic data from many types of experiment.

This thesis has three parts. In the first part, I explore the patterns of indels created by CRISPR-Cas9 genome editing. By thorough characterization of the precision of editing at thousands of genomic target sites, we identify simple sequence rules that can help predict these outcomes. Furthermore, we examine the role of the structural chromatin context in fine-tuning Cas9-DNA interactions.

In the second part, I explore methods to study protein-RNA interactions. I use comparative computational analyses to assess both the data quality of, and data analysis methods for, different crosslinking and immunoprecipitation (CLIP) technologies. I then develop new methods to analyse data generated by hybrid individual-nucleotide resolution CLIP (hiCLIP). By tailoring computational solutions to an understanding of experimental conditions, I improve the overall sensitivity of hiCLIP, and ultimately feedback to drive ongoing experimental development.

In the third part, I focus on the Staufen family of double-stranded RNA binding proteins and using hiCLIP data to define transcriptome-wide atlases of RNA duplexes bound by these proteins both in a cell line and in rat brain tissue. Through integration with other data sets, both publicly available and newly generated, I derive insights into their function in RNA metabolism, and in how these interactions change during the course of mammalian brain development with putative roles in ribonucleoprotein complex formation.

In summary, I present a range of tailored computational methods and analyses developed to understand interactions between proteins and nucleic acids; aiming to link these interactions to functional outcomes.
IMPACT STATEMENT

The work presented in this thesis contributes broadly to our understanding of how proteins and DNA or RNA molecules interact. We have derived both new biological insights and developed new computational methods that have advanced our understanding and also provides new tools and frameworks to help ongoing research in these fields.

Our work on CRISPR-Cas9 technology discovered a set of simple rules that determine the precision of gene editing outcomes. Previously, although it had been discovered that the patterns of these outcomes were not random, it had been assumed that a range of outcome occurred at a given site. We showed that it was possible to identify sites that reproducibly had the same edit, and defined factors that resulted in this precision. This has important implications for its application both as a research tool and as a future therapeutic intervention. By considering the rules we uncovered when designing guide RNAs that direct the Cas9 enzyme to the target DNA site, we are able to maximise the chance of getting a desired gene editing outcome: either knocking out a gene, or engineering a specific edit to the genetic code. This work has been published, with both raw and processed data and analysis code publicly available.

Protein-RNA interactions are vital to many cellular functions. Mutations in many RNA binding proteins are associated with a number of diseases. In order to understand the disease mechanisms, it is first important to understand the nature of these interactions in health. To do so, we use experimental methods to identify the binding sites of proteins on RNA molecules. I set out methods to gauge the quality of the data generated and compare different computational methods to analyse these data to help guide best practice. The structure of RNA molecules also impacts on the nature of the binding, with some proteins binding double-stranded RNA. I
developed computational methods to improve the yield from existing data, which has also informed improvements in the experimental method. Part of this work has already been published.

Finally, we have used these methods to study the Staufen protein family, which has been implicated in neurodegenerative diseases. Our work generating atlases of Staufen binding on RNA structures, has advanced our understand of both the functions of Staufen and also given us some clues as to how it performs its roles, in particular through assisting ribonucleoprotein complex formation using these structures. Currently, this benefits our understanding of neuronal development and function, but in the future will help contextualise its role in disease.
This thesis describes work carried out at The Francis Crick Institute and University College London between December 2015 and September 2019 under the supervision of Prof. Nicholas Luscombe and Prof. Jernej Ule. The research was supported by a Wellcome Trust Clinical PhD Fellowship.

I performed the majority of the computational analysis, but all of the new experimental work presented here was performed by collaborators as indicated. Work presented in Chapter 2 was a collaboration with the laboratory of Dr Paola Scaffidi (The Francis Crick Institute). The large-scale experiments were undertaken by Mr Tristan Henser-Brownhill, Dr Josep Monserrat and Dr Paola Scaffidi, and the subsequent validation experiments by Dr Josep Monserrat. The neural network analysis was undertaken by Mr Tristan Henser-Brownhill.

The rest of the work was performed with the laboratory of Prof. Jernej Ule as part of a long-standing collaboration between the Luscombe and Ule labs. The microarray data in Chapter 3 was processed by Dr Nejc Haberman. The RNA map script was developed from a version originally written by him. In Chapter 5, rat brain tissue was kindly provided by Dr Sandra Fernández-Moya and Ms Janina Ehses in the laboratory of Prof. Michael Kiebler (Ludwig-Maximilians-Universität, Munich, Germany). The hiCLIP method was optimised and developed for use in tissue by Dr Flora Lee, who performed all of the new hiCLIP and iCLIP experiments described.

All the sequencing was done by the Advanced Sequencing Facility at The Francis Crick Institute.
ACKNOWLEDGEMENTS

During my PhD, I have been very fortunate to meet many interesting people who have all contributed to make the journey the wonderful experience it has been. I would like to acknowledge the Wellcome Trust for funding me throughout my studies. First, I am most indebted to Nick Luscombe for taking a chance on a clinician who had never done any bioinformatics. His support, guidance, encouragement and mentorship have enabled me to develop as a scientist. Second, I am immensely grateful to Jernej Ule for welcoming me into his inspirational research environment. I have benefitted greatly from his enthusiasm, knowledge, guidance and vision. Thank you both for the wide-ranging opportunities you have given me.

I would also like to thank all the members of the Luscombe and Ule labs, who have provided countless hours of enjoyable discussion and debate, both scientific and otherwise. I hope the friendships made long continue. Particular thanks are due to Federico Agostini, who took me under his wing when I was taking my first steps in computational biology, and Anna Poetsch, Nejc Haberman, Cristina Militti and Flora Lee, with whom I worked on the projects I have presented here. I would also like to thank Rolf Backofen, for valuable feedback during his sabattical at The Francis Crick Institute. I would like to apologise and thank in equal measure Jan Attig and Martina Hallegger: for cutting the LINE and TDP-43 projects from my thesis, and for letting me be part of them.

I have learnt much from my collaborators, especially Paola Scaffidi, Tristan Henser-Brownhill and Josep Monserrat. Working together with you was a joy and a privilege and set the benchmark for all my future collaborations.

Lastly, I would like to thank my parents and Jo for their love, encouragement and support in all my endeavours. This thesis is dedicated to you.
## CONTENTS

1 INTRODUCTION .......................................................... 29

1.1 Protein-DNA interactions ........................................... 31
    1.1.1 Histones, chromatin and the packaging of DNA .......... 31
    1.1.2 The histone ‘code’ and transcriptional regulation ......... 33

1.2 Protein-RNA interactions ............................................ 35
    1.2.1 RNA binding proteins as regulators ......................... 36
    1.2.2 The 3′ UTR is an important regulatory hotspot ......... 37
    1.2.3 The importance of post-transcriptional regulation in neurons 40
    1.2.4 The domain structure of RNA binding proteins and their modes of binding ........................................... 43
    1.2.5 Genome-wide methods to study the specificity of protein-RNA interactions ........................................... 45
    1.2.6 RNA binding proteins and human disease ................ 49

1.3 Integration of genomic approaches ............................. 52

1.4 Overview of the thesis .............................................. 53

2 INSIGHTS INTO INTERACTIONS BETWEEN EUKARYOTIC DNA AND CRISPR-CAS9 .................................................. 55

2.1 Summary of the chapter .............................................. 56

2.2 Introduction .......................................................... 56
    2.2.1 What is genome editing? ....................................... 56
    2.2.2 A brief history of CRISPR-Cas9 ............................. 57
    2.2.3 DNA repair pathways .......................................... 61
    2.2.4 Characterising on and off target effects ................. 63
3.2.3 Computational quality control of CLIP data ............... 107
3.2.4 Peak calling approaches for CLIP data ................. 111
3.2.5 An evaluation of CLIP methods and analysis approaches ... 113

3.3 Methods ................................................. 114
3.3.1 Processing of CLIP data .......................... 114
3.3.2 Processing of PTBP1 knockdown splicing microarray data 118
3.3.3 Processing of PTBP1 knockout RNA-seq data ............ 118
3.3.4 Generating RNA maps ............................... 119

3.4 Results .................................................. 120
3.4.1 An overview of the data sets ......................... 120
3.4.2 RNA maps to compare PTBP1 CLIP data ............... 122
3.4.3 Assessing the raw CLIP data quality ................. 123
3.4.4 Assessing the effect of peak calling on data quality .... 125
3.4.5 Assessing the effect of different peak callers on data quality . 127
3.4.6 Assessing optimisation of peak calling parameters ....... 127

3.5 Conclusions ............................................. 132

4 DEVELOPMENT OF METHODS TO STUDY PROTEIN-RNA DUPLEX INTERACTIONS 133

4.1 Summary of the chapter .................................. 134
4.2 The hiCLIP method to study protein-RNA duplex interactions ... 134
4.2.1 The innovations of the hiCLIP method ............... 135
4.2.2 Hypotheses of the sources of ‘lost’ duplexes during data analysis 137

4.3 Methods .................................................. 141
4.3.1 Processing of hiCLIP data .......................... 141
4.3.2 Alignment of non-hybrid reads ....................... 141
4.3.3 Identification of putative duplexes from aligned non-hybrid reads ......................... 142
4.3.4 Identification of putative duplexes protected from RNase digestion ......................... 143
4.3.5 Identification of confident duplexes .................. 144
4.4 Development of improved computational methods for hiCLIP data analysis .............................................. 144
4.4.1 The linker adapter may be partially degraded .......... 144
4.4.2 The linker adapter is not necessary to form hybrid reads .. 144
4.4.3 Short-range duplexes are protected from RNase digestion . 147
4.4.4 Extending the yield of confident hiCLIP duplexes .......... 150
4.5 Conclusions ............................................................. 151

5 INSIGHTS INTO INTERACTIONS BETWEEN RNASTRUCTURES AND STAUEN PROTEINS 153

5.1 Summary of the chapter ................................................. 154
5.2 Introduction .............................................................. 155
5.2.1 The Staufen family of RNA binding proteins .............. 155
5.2.2 Staufen functions: localisation, stability, translation ...... 158
5.2.3 Experimental methods to assess RNA structure .......... 162
5.3 Methods: STAU1 ......................................................... 167
5.3.1 Processing of PARIS data ........................................ 167
5.4 Results: STAU1 ........................................................... 168
5.4.1 Contextualising STAU1-bound duplexes .................... 169
5.4.2 STAU1 binding and RNA metabolism ........................ 173
5.5 Discussion: STAU1 ....................................................... 177
5.5.1 Differences between hiCLIP and PARIS are a combination of biological and technical effects .............. 177
5.5.2 STAU1-bound structures may have roles in RNA metabolism 178
5.6 Methods: STAU2 ........................................................ 179
5.6.1 The hiCLIP experimental method adapted to tissue ....... 179
5.6.2 hiCLIP data pre-processing and read alignment .......... 181
5.6.3 Hybrid read identification and duplex delineation ....... 182
5.6.4 Gene ontology analysis ........................................... 183
5.6.5 Processing of rat brain RNA-seq data ...................... 183
5.6.6 Miscellaneous analyses ........................................... 185
LIST OF FIGURES

1.1 The X-ray crystal structure of the nucleosome core particle . . . . . . 32
1.2 Chromatin organisation in the mammalian nucleus . . . . . . . . . . 33
1.3 Schematic representation of chromatin configurations . . . . . . . . . 34
1.4 Regional binding of RBPs on mRNAs . . . . . . . . . . . . . . . . . . 38
1.5 A schematic of alternative 3′ UTR isoforms . . . . . . . . . . . . . . 40
1.6 Regulation of mRNA in neurons . . . . . . . . . . . . . . . . . . . . . 41
1.7 The domain structure of RBPs . . . . . . . . . . . . . . . . . . . . . . . 44
1.8 Modes of tandem RBD-RNA interactions . . . . . . . . . . . . . . . . 46

2.1 Schematic of the prokaryotic genomic CRISPR locus . . . . . . . . . . 58
2.2 Schematic of the CRISPR-Cas9 complex cleaving DNA . . . . . . . . 60
2.3 Selected non-homologous end joining repair pathways . . . . . . . . 62
2.4 Overview of the experimental setup . . . . . . . . . . . . . . . . . . . 65
2.5 Sequencing library metrics . . . . . . . . . . . . . . . . . . . . . . . . . 71
2.6 The indel identification approach is robust to sequencing errors . . . 72
2.7 Comparison of large-scale and validation experiments . . . . . . . . 74
2.8 Comparison of large-scale and validation experiments . . . . . . . . 74
2.9 The sizes and frame-shifting effects of indels . . . . . . . . . . . . . . 75
2.10 Clustering of indel distributions at each target . . . . . . . . . . . . . 76
2.11 Site-specific precision of editing . . . . . . . . . . . . . . . . . . . . . 77
2.12 Genome browser views of site-specific precision of editing . . . . . 78
2.13 Relationships between editing precision and features of indels . . . 79
2.14 Microhomology-associated deletions . . . . . . . . . . . . . . . . . . . 80
2.15 Microhomology-associated insertions . . . . . . . . . . . . . . . . . . 81
2.16 Relationship between precision and insertion homology . . . . . . . 82
LIST OF FIGURES

2.17 A neural network identifies protospacer nucleotide positions important for editing outcome ........................................ 83
2.18 The nucleotide content of the precision core .................................................. 84
2.19 The nucleotide at position –4 and editing outcome ........................................ 85
2.20 Predicting editing outcome using the precision core ........................................ 86
2.21 Chromatin modulation affects editing efficiency ............................................. 87
2.22 Effect of chromatin modulation on editing outcome at selected targets ............. 88
2.23 Chromatin modulation results in small changes in indel profiles ............... 89
2.24 Chromatin modulation does not affect indel profiles at precise targets .......... 91

3.1 An overview of the iCLIP method and its analysis ....................................... 102
3.2 A comparison of iCLIP, eCLIP and irCLIP experimental methods ............... 105
3.3 Anatomy of an RNA map ................................................................. 110
3.4 Regional distribution of crosslinks ......................................................... 122
3.5 RNA maps of CLIP crosslink sites around regulated exons ....................... 124
3.6 RNA maps of CLIP peak sites around regulated exons .................................. 126
3.7 RNA maps showing the effect of different peak callers .................................. 128
3.8 RNA maps showing the effect of different iCount peak calling parameters .... 130
3.9 RNA maps showing the effect of different Paraclu peak calling parameters ..... 131

4.1 An overview of the steps of the original hiCLIP method ................................. 135
4.2 Sources of ‘lost’ duplexes during data analysis ........................................... 137
4.3 The computational analysis workflow ....................................................... 140
4.4 Schematic of the two types of chimeric read .............................................. 145
4.5 Binding energy of duplexes identified with and without a linker adapter .... 148
4.6 Computational prediction of short-range duplexes in non-hybrid reads .... 149
4.7 An expanded set of confident hiCLIP duplexes ........................................ 151

5.1 STAU1 and STAU2 protein isoforms .................................................... 156
5.2 STAU1 binding to ARF1 ................................................................. 157
<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Validation and comparability of hiCLIP and PARIS duplexes</td>
<td>170</td>
</tr>
<tr>
<td>5.4</td>
<td>Binding energy of duplexes</td>
<td>172</td>
</tr>
<tr>
<td>5.5</td>
<td>Regional distribution and span of duplex arms</td>
<td>173</td>
</tr>
<tr>
<td>5.6</td>
<td>Expression levels of STAU1-bound transcripts</td>
<td>174</td>
</tr>
<tr>
<td>5.7</td>
<td>RNA metabolism rates of STAU1-bound transcripts</td>
<td>175</td>
</tr>
<tr>
<td>5.8</td>
<td>3' UTR length and RNA degradation rates</td>
<td>175</td>
</tr>
<tr>
<td>5.9</td>
<td>PolyA sites around duplex arms</td>
<td>176</td>
</tr>
<tr>
<td>5.10</td>
<td>Circularisation score and RNA degradation rates</td>
<td>177</td>
</tr>
<tr>
<td>5.11</td>
<td>The hiCLIP experimental method in rat brain</td>
<td>180</td>
</tr>
<tr>
<td>5.12</td>
<td>Extension of the <em>Rattus norvegicus</em> 3' UTR annotation</td>
<td>187</td>
</tr>
<tr>
<td>5.13</td>
<td>Regional distribution of STAU2 binding</td>
<td>189</td>
</tr>
<tr>
<td>5.14</td>
<td>Principal components analysis of iCLIP binding for STAU2 and mock IP</td>
<td>190</td>
</tr>
<tr>
<td>5.15</td>
<td>STAU2 binding enrichment on mature mRNAs</td>
<td>191</td>
</tr>
<tr>
<td>5.16</td>
<td>STAU2-bound duplexes on the Camk2a 3' UTR</td>
<td>193</td>
</tr>
<tr>
<td>5.17</td>
<td>Types of hybrids detected</td>
<td>194</td>
</tr>
<tr>
<td>5.18</td>
<td>Regions linked by hybrids</td>
<td>195</td>
</tr>
<tr>
<td>5.19</td>
<td>Intragenic duplexes in mature mRNA</td>
<td>196</td>
</tr>
<tr>
<td>5.20</td>
<td>The canonical polyA signal around STAU2-bound 3' duplexes</td>
<td>197</td>
</tr>
<tr>
<td>5.21</td>
<td>The validity of detected duplexes</td>
<td>198</td>
</tr>
<tr>
<td>5.22</td>
<td>Conservation around STAU2-bound duplexes</td>
<td>199</td>
</tr>
<tr>
<td>5.23</td>
<td>Examples of STAU2 binding to RNA duplexes across development</td>
<td>200</td>
</tr>
<tr>
<td>5.24</td>
<td>Principal components analysis of crosslink-quantified STAU2-bound duplexes</td>
<td>201</td>
</tr>
<tr>
<td>5.25</td>
<td>Expression-normalised differential STAU2 duplex binding</td>
<td>202</td>
</tr>
<tr>
<td>5.26</td>
<td>Heatmaps of STAU2 duplex binding at different developmental ages</td>
<td>203</td>
</tr>
<tr>
<td>5.27</td>
<td>Structures of STAU2-bound duplexes</td>
<td>204</td>
</tr>
<tr>
<td>5.28</td>
<td>Regional distribution of HuR and Rbfox1 binding</td>
<td>206</td>
</tr>
<tr>
<td>5.29</td>
<td>HuR and Rbfox1 binding around Camk2a STAU2-bound duplexes</td>
<td>207</td>
</tr>
<tr>
<td>5.30</td>
<td>HuR binding around STAU2-bound duplexes</td>
<td>208</td>
</tr>
<tr>
<td>5.31</td>
<td>Rbfox1 binding around STAU2-bound duplexes</td>
<td>209</td>
</tr>
</tbody>
</table>
LIST OF TABLES

3.1 The central features of CLIP from the perspective of data analysis . . 103
3.2 Overview of PTBP1 CLIP data sets . . . . . . . . . . . . . . . . . 120

5.1 Stages of post-partum rat development . . . . . . . . . . . . . . . 186
5.2 Overview of STAU2 and mock IP iCLIP experiments in rat brain . . 188
5.3 Overview of STAU2 hiCLIP experiments in rat brain . . . . . . . . . 192
5.4 Overview of HuR and Rbfox1 iCLIP experiments in rat brain . . . . 205
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4sU</td>
<td>4-thiouridine</td>
</tr>
<tr>
<td>AGO</td>
<td>Argonaute</td>
</tr>
<tr>
<td>AMT</td>
<td>4′-aminomethyltrioxsalen</td>
</tr>
<tr>
<td>BDNF</td>
<td>brain-derived neurotrophic factor</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>Camk2a</td>
<td>Calcium/Calmodulin dependent protein kinase II alpha</td>
</tr>
<tr>
<td>Cas</td>
<td>CRISPR associated</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>chromatin immunoprecipitation sequencing</td>
</tr>
<tr>
<td>CIRS</td>
<td>chemical interference of RNA structures</td>
</tr>
<tr>
<td>CLASH</td>
<td>crosslinking, ligation and sequencing of hybrids</td>
</tr>
<tr>
<td>CLIP</td>
<td>UV crosslinking and immunoprecipitation</td>
</tr>
<tr>
<td>COMRADES</td>
<td>crosslinking of matched RNAs and deep sequencing</td>
</tr>
<tr>
<td>CPSF</td>
<td>cleavage and polyadenylation stimulating factor</td>
</tr>
<tr>
<td>CRAC</td>
<td>cross-linking and analysis of cDNAs</td>
</tr>
<tr>
<td>CRISPR</td>
<td>clustered regularly interspersed palindromic repeat</td>
</tr>
<tr>
<td>crRNA</td>
<td>CRISPR RNA</td>
</tr>
<tr>
<td>CstF</td>
<td>cleavage stimulating factor</td>
</tr>
<tr>
<td>DMS</td>
<td>dimethyl sulphate</td>
</tr>
<tr>
<td>DNase-seq</td>
<td>DNase I hypersensitive sites sequencing</td>
</tr>
<tr>
<td>DSB</td>
<td>double-strand breaks</td>
</tr>
<tr>
<td>dsRBD</td>
<td>double-stranded RNA binding domain</td>
</tr>
<tr>
<td>dsRBP</td>
<td>double-stranded RNA binding protein</td>
</tr>
<tr>
<td>eCLIP</td>
<td>enhanced CLIP</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>hiCLIP</td>
<td>hybrid iCLIP</td>
</tr>
<tr>
<td>HITS-CLIP</td>
<td>high throughput sequencing CLIP</td>
</tr>
<tr>
<td>hnRNP</td>
<td>heterologous nuclear ribonucleoproteins</td>
</tr>
<tr>
<td>HR</td>
<td>homologous recombination</td>
</tr>
<tr>
<td>HuR</td>
<td>Hu-antigen R</td>
</tr>
<tr>
<td>iCLIP</td>
<td>individual-nucleotide resolution CLIP</td>
</tr>
<tr>
<td>icSHAPE</td>
<td>in vivo click selective SHAPE</td>
</tr>
<tr>
<td>IGF2BP</td>
<td>insulin-like growth factor 2 mRNA binding protein</td>
</tr>
<tr>
<td>Impa1</td>
<td>inositol monophosphatase 1</td>
</tr>
<tr>
<td>IP</td>
<td>immunoprecipitation</td>
</tr>
<tr>
<td>irCLIP</td>
<td>infrared CLIP</td>
</tr>
<tr>
<td>kb</td>
<td>kilobases</td>
</tr>
<tr>
<td>KH</td>
<td>K-homology</td>
</tr>
<tr>
<td>LIGR-seq</td>
<td>ligation of interacting RNA and high-throughput sequencing</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long non-coding RNA</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>MMEJ</td>
<td>microhomology-mediated end joining</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>NHEJ</td>
<td>non-homologous end joining</td>
</tr>
<tr>
<td>NMD</td>
<td>nonsense mediated decay</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotides</td>
</tr>
<tr>
<td>PAM</td>
<td>protospacer adjacent motif</td>
</tr>
<tr>
<td>PAR-CLIP</td>
<td>photoactivateable ribonucleoside-enhanced CLIP</td>
</tr>
<tr>
<td>PARIS</td>
<td>psoralen analysis of RNA interactions and structures</td>
</tr>
<tr>
<td>PARS</td>
<td>parallel analysis of RNA structure</td>
</tr>
<tr>
<td>polyA</td>
<td>polyadenylation</td>
</tr>
<tr>
<td>PTBP1</td>
<td>polypyrimidine tract binding protein 1</td>
</tr>
<tr>
<td>RBD</td>
<td>RND binding domain</td>
</tr>
<tr>
<td>Rbfox</td>
<td>RNA Binding Fox-1 Homolog</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA binding protein</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS

RGG    Arg-Gly-Gly
RNP    ribonucleoprotein
RRM    RNA recognition motif
SDS-PAGE sodium dodecyl sulphate polyacrylamide gel electrophoresis
SELEX systematic evolution of ligands by exponential enrichment
sgRNA single guide RNA
SHAPE selective 2'-hydroxyl acylation and primer extension
SINE short interspersed nuclear element
SMD    Staufen mediated decay
snRNA small nuclear RNA
SPLASH sequencing of psoralen crosslinked, ligated and selected hybrids
ssRBP single-stranded RNA binding protein
STAU1  Staufen1
STAU2  Staufen2
TALE   transcription activator-like effectors
TDP-43 TAR DNA-binding protein 43
tracrRNA trans-activating crRNA
TSA    Trichostatin A
UPF1   upstream frameshift 1
UTR    untranslated region
ZF     zinc finger
INTRODUCTION

1.1 Protein-DNA interactions .................................................. 31
   1.1.1 Histones, chromatin and the packaging of DNA ............... 31
   1.1.2 The histone ‘code’ and transcriptional regulation .......... 33
1.2 Protein-RNA interactions ..................................................... 35
   1.2.1 RNA binding proteins as regulators ............................... 36
   1.2.2 The 3′ UTR is an important regulatory hotspot ............... 37
   1.2.3 The importance of post-transcriptional regulation in neurons ... 40
   1.2.4 The domain structure of RNA binding proteins and their modes of binding ... 43
   1.2.5 Genome-wide methods to study the specificity of protein-RNA interactions ... 45
   1.2.6 RNA binding proteins and human disease ....................... 49
1.3 Integration of genomic approaches .................................... 52
1.4 Overview of the thesis ...................................................... 53
Regulation is fundamental to orchestrating the cellular life-cycle. The central ‘dogma’ focuses on the transfer of sequence information from DNA, to RNA, to proteins; “it says nothing about the control mechanisms” (Crick, 1970). But how much, if any, of the sequence information should be transferred? When should it be transferred? Where should it be transferred? Arguably as important as the content of the information being transferred are the answers to these questions. Within them is held the understanding of how the remarkable diversity of multicellular organisms, both at the level of the individual and of the population, is achieved and coordinated. What makes a neuron a neuron? What makes a healthy neuron a sick neuron? The key is largely held in the control of the information transfer.

Proteins have long been considered uniquely important in biology, with their ability to perform a widely diverse range of functions, from structural to enzymatic (Crick, 1958). Indeed, behind the argument of the central dogma is the premise that the main role of the nucleic acids, DNA and RNA, are to result in the synthesis of proteins, which are the main executors of the genetic code; although, we now know that some non-coding RNA can also act as these executors. In brief, there are two main processes that govern the transfer of information. First, DNA is transcribed into RNA by RNA polymerase (a protein). For some non-coding RNAs, for example microRNAs (miRNAs) or long non-coding RNAs (lncRNAs), this is the end of the transfer: they are effectors themselves. For others, the messenger RNA (mRNA) that contains protein coding information is, in turn, translated into a protein by the ribosome (a complex of proteins and RNA). So, here immediately are two levels at which regulation can occur: transcriptional, or post-transcriptional.

There is remarkably little difference in the number of protein coding genes between mammals and worms: the size of the protein coding genome plateaus with increasing organismal complexity (Liu et al., 2013). Therefore, pivotal to the development of the complexity of higher organisms is its regulation. Although the transfer of genetic information is largely unidirectional, the regulatory interactions controlling it are both multifaceted and multidirectional, through often complex regulatory networks. One major set of regulatory interactions is between proteins and nucleic acids.
Here, I start by introducing some examples of protein-DNA and protein-RNA interactions that are pivotal to proper cellular function. Then, I consider the importance of integrating different types of genomic data to gain meaningful biological insight into these interactions.

1.1 PROTEIN-DNA INTERACTIONS

In terms of the regulatory role of protein-DNA interactions, there are two major groups that ensure correct control of transcription and coordination of gene expression: i) transcription factors and ii) histones. Transcription factors have been defined as a class of proteins that bind DNA in a sequence-specific manner, and can regulate transcription (Lambert et al., 2018; Vaquerizas et al., 2009). In the latest cataloguing of the human transcription factor repertoire, 1,639 known or likely such proteins were characterised (Lambert et al., 2018). Transcription factors have a wide range of effector functions upon binding DNA, ranging from directly recruiting RNA polymerase, to steric hindrance of the binding of other proteins, but largely their main mode of action is through the recruitment of co-factor protein complexes that act as co-activators or co-repressors of transcription (Lambert et al., 2018). Study of these protein-DNA interactions are not within the scope of my thesis.

1.1.1 Histones, chromatin and the packaging of DNA

Histones were first described by Albrecht Kossel in 1884 (Campos and Reinberg, 2009; Olins and Olins, 2003) and their association with nucleic acids uncovered in his work¹. They are remarkably invariant, highly conserved proteins with only five histone types: H1/H5, H2A, H2B, H3 and H4 (Kornberg and Lorch, 1999; Phillips and Johns, 1965). They are also small proteins, with molecular weights of approximately 11–13 kDa.

Interactions between histone proteins and DNA, and between proteins and

¹Indeed, in eukaryotic chromosomes they are found in equal levels to DNA, and were long-viewed as the source of genetic material (Kornberg and Lorch, 1999)
proteins result in the formation of higher-order structures. The functional unit of this organisation is the nucleosome. On the basis of biochemical studies, this was predicted to be, “a repeating unit of eight histone molecules and about 200 DNA base pairs” (Kornberg, 1974). This octamer was formed of two each of H2A, H2B, H3 and H4, with H1 proposed to bind outside the nucleosome (Kornberg, 1974; Kornberg and Lorch, 1999). The later X-ray crystal structure of the nucleosome supported this model (Figure 1.1), and in addition showed how the histone octamer was assembled, with 145–147 base pairs (bp) of DNA wrapped around it in a superhelix with 1.65 turns (Luger et al., 1997). Importantly, histones also have unstructured N-terminal tails that extend out of the core nucleosome and can be post-translationally modified. Individual nucleosomes are bridged by the linker histone, H1, and variable lengths of linker DNA. This forms the basic unit of chromatin, a structure first visualised by Walter Flemming in 1882 (Flemming, 1882; Olins and Olins, 2003) and that has further higher order structures (Figure 1.2).
1.1. PROTEIN-DNA INTERACTIONS

Figure 1.2: Chromatin organisation in the mammalian nucleus
(A) Chromosomes are organised in chromosome territories.
(B) Chromosome territories are comprised of fractal globules.
(C) Chromatin fibres interact within globules or territories.
(D) Chromatin may form a 30 nm fibre with a solenoid, zigzag, or polymer melt organisation.
(E) Chromatin is resolved as a 10 nm ‘beads on a string’ fibre consisting of nucleosomes. Reprinted from Hübner et al. (2013), with permission from Elsevier.

1.1.2 The histone ‘code’ and transcriptional regulation

For many years since the identification of DNA as the source of the genetic code histones were relegated to a role as scaffolding, only required for the compaction of DNA into the nucleus. However, it has become apparent chromatin structure plays important roles in many DNA-related processes, including the regulation of gene transcription.

Central to these functions is the N-terminal tail and its post-translational modifications. There are hundreds of such modifications on specific amino acid positions across eight classes, including acetylation and methylation, that each have different overlapping regulatory functions (Kouzarides, 2007). Histone modification, together with DNA methylation, form the two cornerstones of epigenetic regulation. There are three, not mutually exclusive, proposed mechanisms by which they can effect their function: i) they can loosen or tighten inter- or intra-nucleosomal histone-DNA interactions; ii) different modifications can be recognised by different effector non-histone proteins that can direct regulation; and iii) they can directly influence higher order chromatin structure (Kouzarides, 2007; Li et al., 2007). Furthermore, these modifications are highly abundant on the histone tail, which makes crosstalk and co-operative or antagonistic effects likely (Kouzarides, 2007).
Chapter 1. Introduction

Figure 1.3: Schematic representation of chromatin configurations
Models for euchromatic or heterochromatic histone tail modifications. From Jenuwein and Allis (2001). Reprinted with permission from AAAS.

As a model to interpret the epigenetic effects of these modifications, the concept of a histone ‘code’ has been proposed as a means of extending the information contained in the genetic code by regulating access to it (Jenuwein and Allis, 2001). This suggests two overarching states of chromatin: euchromatin, where DNA is accessible to the transcription machinery, and heterochromatin, where it is inaccessible: on or off states (Figure 1.3). The state of the chromatin is dependent on the combinations of histone modifications that are present (Jenuwein and Allis, 2001; Turner, 2000). A two-state model, although an important starting framework, is too simplistic to capture the range of configurations. More recent computational analysis of large genomic data sets have surveyed this landscape of epigenetic modifications at scale and permitted finer resolution of these two states into 8 active and 7 repressed states, with regions corresponding to, amongst others, enhancers, promoters, active transcriptional start sites and constitutive heterochromatin, for example (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium et al., 2015). Ultimately, there is likely to be a continuum of accessibility.

I briefly expand on two families of modification relevant for this thesis, acetylation and methylation. Acetylation is almost always associated with activation of transcription. This is likely because it neutralises the basic charge of the lysine and so has the most potential to unfold chromatin, rendering it accessible (Kouzarides, 2007). Moreover, the acetyltransferase family responsible is relatively non-specific and generally acetylates more than one lysine residue. Reciprocally, deacetylation, by families of histone deacetylases, is associated with transcriptional repression. Methylation carried out by lysine methyltransferases, on the other hand, is highly
1.2. PROTEIN-RNA INTERACTIONS

specific and can modify individual lysines on single histones (Kouzarides, 2007). As such they can have activating or repressing roles dependant on the residue modified.

Finally, as a prelude to Chapter 2, I note that while these interactions may regulate access to the genetic code, they by and large do not change it. However, we are now introducing further complexity into biological systems with our drive to alter the underlying genetic code: either as a scientific tool to investigate biological systems, or as potential therapeutic agents for genetic diseases (Park et al., 2016b). Hence, also understanding how exogenous factors may affect these protein-DNA interactions is an important consideration for the future.

1.2 PROTEIN-RNA INTERACTIONS

Rather than a simple transitional molecule between the two ‘important’ ends of the central dogma model, there is much evidence to place RNA at its centre (Sharp, 2009). It is able to catalyse reactions (commonly considered protein functions). It is able to encode information (commonly considered DNA functions). Indeed the RNA world hypothesis proposes that it may have been the original molecule of life (Gilbert, 1986). While these functions have been expropriated by protein and DNA, RNA remains more than a mere “infrastructural platform” for protein synthesis, but has important and wide-ranging regulatory roles (Morris and Mattick, 2014; Sharp, 2009). Exploring the latest reference annotation of the human genome (Frankish et al., 2019), shows that of the 60,303 annotated genes, 19,975 are annotated as protein coding, with 16,480 as IncRNA, and 1,881 as miRNA. While RNA can interact with other RNA, DNA and protein molecules, here I focus on the latter.

With the discovery of splicing it became apparent that regulation of RNA could be another key mechanism to control the flow of information from DNA to protein. Moreover, the finding that through splicing, different transcripts containing different combinations of exons could be created from the same pre-mRNA, means that RNA regulation has the possibility to generate the complexity seen in higher organisms in the face of only a modest increase in protein coding genes. Indeed studies assessing

2GENCODE Human Release 31 (GRCh38.p12)
alternative splicing across a range of human tissues have found that 92–95% of genes undergo alternative splicing (Pan et al., 2008; Wang et al., 2008) and importantly these were not rare events with ∼86% with a minor isoform frequency of over 15% (Wang et al., 2008).

1.2.1 RNA binding proteins as regulators

Early studies found that RNA binding proteins (RBPs), such as the heterologous nuclear ribonucleoproteins (hnRNPs), were important for the regulation of the splicing process. This led to the discovery of functional RNA binding domains (RBDs) contained within the proteins (Burd and Dreyfuss, 1994): most prevalent being the RNA recognition motif (RRM), with others such as the K-homology (KH) domain, zinc finger domain, Arg-Gly-Gly (RGG) box, Puf binding repeats, and double-stranded RNA binding domains (dsRBDs) (Glisovic et al., 2008; Lunde et al., 2007).

A number of proteomic and computational studies have attempted to quantify the RNA binding proteome; that is how many RBPs are there? Estimates vary, ranging from 1,207 to 1,838 RBPs (Castello et al., 2012; Caudron-Herger et al., 2019; Gerstberger et al., 2014; Hentze et al., 2018; Queiroz et al., 2019; Trendel et al., 2019), with the some of the higher estimates extending the concept of ‘binding’ from a binary state to consider proteins that may be more loosely or transiently associated with RNAs, but still dependent on these interactions for their action. Indeed many newly discovered RBPs lack canonical RNA binding domains, but in fact DNA-binding domains, protein-protein interaction interfaces and intrinsically disordered regions have been found to be involved in associations with RNA (Hentze et al., 2018). So, with potentially thousands of RBPs and tens of thousands of RNAs, there is significant scope for post-transcriptional regulation of RNAs by proteins. Moreover ∼98% of these are not tissue-specific (Gerstberger et al., 2014), suggesting they have general, conserved cellular functions.

Interactions between RBPs and non-coding RNAs have important regulatory
1.2. PROTEIN-RNA INTERACTIONS

roles. For example, the lncRNA NEAT1 forms a scaffold upon which certain RBPs can bind to form paraspeckles, which are important in processes such as embryonic differentiation (Modic et al., 2019). Or miRNAs require interactions with the RBPs Drosha and Dicer for their biogenesis and maturation and together with the protein Argonaute (AGO) form a silencing complex to fulfil their effector function (Bartel, 2018). In my thesis, however, I focus primarily on interactions with the protein coding transcriptome.

mRNA is in complex with a diverse and varying complement of RBPs throughout its life-cycle, from the transcription nascent pre-mRNA transcript in the nucleus till the degradation and decay of the mature mRNA in the cytoplasm (Moore, 2005; Singh et al., 2015). These make up the messenger ribonucleoprotein (mRNP). If we consider an RNA-centric perspective, the mRNP is assembled immediately co-transcriptionally with the subsequent addition or removal of RBPs as necessary for aspects of mRNA processing and life: 5’ end capping, splicing, 3’ end cleavage and polyadenylation (polyA) in the nucleus during the maturation of the pre-mRNA, transport through the nuclear pore, localisation within the cell, translation and finally degradation (Moore, 2005). Furthermore RBPs bind throughout the mRNA molecule (Figure 1.4) with regional distributions that correspond to distinct functional roles of the RBP.

1.2.2 The 3’ UTR is an important regulatory hotspot

Here, I focus on the 3’ untranslated region (3’ UTR) of mRNA transcripts to explore how protein-RNA interactions have a diverse range of effects not just on the RNA, but also on the protein.

The 3’ UTR spans the region between the stop codon and the polyadenylated tail of the mRNA. This region of the mRNA transcript has lengthened considerably during evolution: from a median of 140 nucleotides (nt) in worms to 1,200 nt in humans (Mayr, 2016). This allows for the possibility of greater post-transcriptional gene regulation. Two other observations support the premise that the 3’ UTR has an
Figure 1.4: Regional binding of RBPs on mRNAs
The distribution of binding of RBPs on regions of the mRNA transcript. Raw data obtained from the ENCODE consortium (Van Nostrand et al., 2017a) and processed using methods described in Chapter 3. RBPs are clustered hierarchically by their binding preferences (dendrogram not displayed).
important role in the complexity of higher organisms: i) an expansion in genome size in multicellular organisms correlates with this increase in length (Mayr, 2017); and ii) there is an exponential correlation between 3′ UTR length and morphological complexity (as measured by number of cell types) (Chen et al., 2012; Mayr, 2017). It is therefore likely that the 3′ UTR contributes to the evolution of a complex functional proteome in humans in the face of a relatively constant number of protein-encoding genes discussed earlier.

As can be seen in Figure 1.4 there are a subset of RBPs that preferentially bind to the 3′ UTR. This 3′ UTR “compositional binding domain” (Singh et al., 2015) is enriched for proteins involved in polyadenylation, localisation, translation and decay, such as cleavage and polyadenylation stimulating factors (CPSFs), TAR DNA-binding protein 43 (TDP-43), Insulin-like growth factor 2 mRNA binding proteins, also known as zipcode binding proteins (IGF2BPs), Staufen, Hu-antigen R (HuR), AGO and UPF1. Indeed, these trans-acting factors are involved in the gamut of known 3′ UTR functions in the RNA life-cycle (Mayr, 2017). Moreover, the interactions not only have effects on the bound RNA, but there is also evidence that the RNA can affect how the protein can function, particularly through the formation of complexes. The 3′ UTR can act as a scaffold for proteins and thus regulate the interactors of the protein encoded for by the transcript (Berkovits and Mayr, 2015; Lee and Mayr, 2019; Ma and Mayr, 2018; Mayr, 2017). Alternative 3′ UTR isoforms thus enable the formation of the same protein, but in different protein complexes and cellular locations (Berkovits and Mayr, 2015; Mayr, 2017).

Part of the complexity generated by 3′ UTRs comes from variability in their length for a given gene. Analogous to alternative splicing, which can generate alternative protein-coding mRNA transcript isoforms, alternative cleavage and polyadenylation can generate 3′ UTR isoforms of different lengths for a given mRNA transcript by the selection of alternative polyA sites. This leads to transcripts where longest and shortest isoforms have a seven-fold difference in median length — 1,773 nt and 249 nt for example in mice (Tian and Manley, 2017). Between 51–79% of genes

---

3 Alternative splicing can also contribute to 3′ UTR isoform diversity through the use of alternative last exons (Taliaferro et al., 2016)
produce different 3’ UTR isoforms (Derti et al., 2012; Lianoglou et al., 2013; Mayr, 2017). Moreover, there are tissue-specific preferences for isoforms of different length (Derti et al., 2012; Lianoglou et al., 2013). In brain tissue there is a general tendency towards distal polyA site usage, resulting in longer 3’ UTR isoforms (Lianoglou et al., 2013; Miura et al., 2013; Zhang et al., 2005). Indeed in a number of cases neuronal 3’ UTRs with lengths of 10–20 kilobases (kb) have been observed (Miura et al., 2013). Consequently the presence of an additional RNA segment enables greater regulatory diversity (Figure 1.5).

1.2.3 The importance of post-transcriptional regulation in neurons

The usage of longer 3’ UTRs in neurons means that there is the possibility for more regulatory elements, such as RBP binding sites. Interestingly within these extended alternative 3’ UTR regions, over-represented sequences have been found that likely correspond to the binding sites of particular RBPs (Tushev et al., 2018; Zhang et al., 2005). While a number of protein-RNA interactions have been found to regulate polyA site selection, including the binding of proteins such as NOVA (Licatalosi et al., 2008), SRSF3 (Müller-McNicoll et al., 2016) and TDP-43 (Rot et al., 2017),
1.2. PROTEIN-RNA INTERACTIONS

Figure 1.6: Regulation of mRNA in neurons
Reprinted from Glock et al. (2017), with permission from Elsevier.

rather than the role of protein-RNA interactions in regulating 3′ UTR isoforms, here I discuss their roles within the functions of the longer 3′ UTR isoforms, with a particular focus on RNA localisation and stability.

The appropriate cellular localisation of RNAs is important; in one study in *Drosophila*, 71% of analysed genes that were expressed encoded transcripts with subcellular localisation patterns (Lécuyer et al., 2007). Neurons are highly polarised cells with dendrites up to tens of millimetres long, and axons up to a metre; the soma may only contain 1% of the cytoplasm (Holt et al., 2019). This unique degree of subcellular compartmentalisation has resulted in the necessity to transport mRNA from the nucleus to the ends of neurites and there maintain a local milieu of RNA and protein to ensure a rapid local response to stimuli through local translation of mRNA (Andreassi and Riccio, 2009; Biever et al., 2019b; Glock et al., 2017; Hafner et al., 2019; Holt et al., 2019) as summarised in Figure 1.6.

In response to long-term potentiation induction, newly synthesised Calcium/Calmodulin dependent protein kinase II alpha (Camk2a) protein can be detected at the dendritic synapse within 5 minutes — a speed that cannot be accomplished by macromolecular transport mechanisms (Ouyang et al., 1999). Hence, rather than somatic translation with protein transport to the synapse, the mRNA is transported to the synapse, where there is also located a pool of translation machin-
This enables both spatial restriction of gene expression and also a high temporal resolution: stimulation of the synapse can enable the local translation of an mRNA pool that is already present, rather than requiring a signal to be transmitted back to the nucleus in the soma, mRNA transcribed, exported, translated in the cytoplasm and then transported down to the synapse. Moreover, local translation is also energetically efficient as multiple copies of a protein can be translated on demand from one already-transported mRNA molecule (Martin and Ephrussi, 2009). This means that there can be a fast response to localised cues, circumventing long transport times. A range of proteins have been found to be locally translated to fulfil many functions including synapse formation, axon branching and maintenance and synaptic plasticity (Holt et al., 2019). Regulation of both mRNA localisation and stability by 3′ UTRs to maintain these localised mRNA colonies are therefore particularly important in neurons, and this may account for the observed 3′ UTR lengthening.

Trafficking of mRNA occurs in a microtubule-dependant manner: RBPs bind localisation elements in the 3′ UTR, and then in turn bind motor proteins (either directly, or through other protein-protein interactions) that enable the mRNP to move across the cell (Holt et al., 2019; Martin and Ephrussi, 2009; Mayr, 2017). There are well characterised cases of alternative 3′ UTR usage leading to differential localisation of the short and long isoforms to either the soma or the neurite: for example, the brain-derived neurotrophic factor (BDNF) mRNA (An et al., 2008) or the inositol monophosphatase 1 (Impa1) mRNA (Andreassi et al., 2017). Genome-wide assessment of different 3′ UTR isoforms in hippocampal neurons found transcripts with multiple 3′ UTR isoforms to have different subcellular localisation (Tushev et al., 2018). Localisation elements, recognised by trans-acting RBPs, have been characterised in neurons for a number of mRNA transcripts and are usually located in the 3′ UTR (Andreassi and Riccio, 2009).

In neurons, transcripts with the longer 3′ UTR isoforms found in the neuropil were found to have longer half-lives than their shorter 3′ UTR isoforms in the soma (Tushev et al., 2018). 3′ UTRs are also important for the regulation of mRNA stability more generally. As mentioned, interactions between the 3′ UTR and an RNP
1.2. PROTEIN-RNA INTERACTIONS

consisting of a miRNA and AGO are important to direct the post-transcriptional repression of mRNAs (Bartel, 2009). Hotspots for regulatory binding of other trans-acting factors including RBPs, that are involved in RNA stability in 3′ UTRs have also been described (Plass et al., 2017). Indeed, the nonsense mediated decay pathway, which selectively degrades mRNAs as a means of post-transcriptional regulation, is mediated through the ATP-dependent RNA helicase upstream frameshift 1 (UPF1) and its binding to 3′ UTR elements (Kim and Maquat, 2019).

1.2.4 The domain structure of RNA binding proteins and their modes of binding

So far, I have given an overview of the functional implications of protein-RNA interactions, with a particular focus on the 3′ UTR and neurons. To do so, I have largely given non-specific descriptions of interactions between protein ‘domains’ and RNA ‘sites’. However, understanding the RNA binding specificities of RBPs is critical to elucidating the mechanisms behind post-transcription regulation by their interactions. Although both RNA sequence and RNA structure are likely to be important in directly specificity, most work to date has been on sequence. One of the first 3′ UTR regulatory sequences identified was a highly conserved “AAUAAA” hexamer 20 nt upstream of the 3′ polyA site and predicted to be a signal for cleavage and polyadenylation of the mRNA transcript (Proudfoot and Brownlee, 1976). Further biochemical studies identified the link between the signal and the process as the recognition of the signal by protein complexes CPSF and CstF (cleavage stimulating factor) and their subsequent binding (Proudfoot, 2011).

As for many proteins, RBPs have a modular structure (Gallagher and Ramos, 2018; Gerstberger et al., 2014; Lunde et al., 2007). Moreover, they are often comprised by repeating the basic binding domains introduced earlier (Figure 1.7). The existence of multiple RBDs opens the door to multiple modes of RNA binding by RBPs through combinations of domains acting synergistically (or indeed antagonistically in autoregulatory roles). This is important as structural studies have shown that a single RBD usually only contacts a 4 nt stretch in the RNA molecule (Mackereth
Figure 1.7: The domain structure of RBPs
Reprinted by permission from Springer Nature: Lunde et al. (2007).

and Sattler, 2012): this both limits the potential specificity of a given RBP and also often results in binding with weak affinity (Lunde et al., 2007). Hence, by combining multiple RBDs it is possible to increase both specificity and affinity of an RBP through repeated low affinity or weak interactions. Combinatorial binding also enables multiple functional roles to be fulfilled by an RBP depending on the combination of RBDs in an RBP or indeed RBPs in an RNP.

One example of multiple domains conferring target specificity is the RBP Zipcode-binding protein 1 (ZBP1) which has two RRM domains and four KH domains. Depending on the target RNA, either two (KH1-KH2) or four (KH1-KH2-KH3-KH4) domains are needed to enable recognition and appropriate binding (Gallagher and Ramos, 2018). Nuclear magnetic resonance studies have shown that when ZBP1 binds to the β-actin mRNA the KH3 domain binds to an ACAC motif while the KH4 domain binds to a CGGAC motif and that these binding sites
are on opposite sides of KH3-KH4 domain complex (Gallagher and Ramos, 2018; Nicastro et al., 2017). The coupled domains bind with a 50-fold higher affinity than the individual domains. This example also highlights the observation that different KH domains can have different sequence specificities or binding affinities for a given sequence: KH4 is highly specific for a longer, more complex motif, while KH3 is less specific, recognising a shorter more degenerate motif. One speculation is that the KH3 domain may be scanning the mRNA to ‘short-list’ binding sites which are confirmed by KH4 binding (Nicastro et al., 2017).

These interactions also scale to RNPs, where multiple RBPs act in concert to effect a particular function. Hrp1 and Rna15 proteins are important in 3′ end processing of pre-mRNAs. However, in isolation, the RRM of Rna15 has little binding specificity or affinity; it is when the two RRMs of Hrp1 bind to the adjacent enhancer element that high affinity binding of Rna15 to the AU-rich positioning element is detected (Mackereth and Sattler, 2012).

There are many such examples of the RBDs being combined in different ways to perform different functional roles. These approaches are summarised in Figure 1.8. In addition to acting independently (Figure 1.8B), RBDs can act together to facilitate high-affinity specific binding through conformational changes in the RBP (Figures 1.8C, D) and through directing conformational changes in the RNA (Figure 1.8E, e.g. PTBP1). Furthermore, auxiliary protein-protein contacts either from a different protein (Figure 1.8F), or through dimerisation of the same protein using a different domain (Figure 1.8G, e.g. STAUI1) can also enhance binding affinities.

1.2.5 Genome-wide methods to study the specificity of protein-RNA interactions

Much insight has been gained from resolving structures of RBPs, or particular RBDs within an RBP, bound to its target mRNA sequence. However, these are individual examples of binding for a given RBP; integrating these results with large-scale genome-wide methods is necessary to complement these insights to derive generalisable principles of the mechanisms of protein-RNA interactions.
Both in vitro and in vivo approaches have been taken to address the specificity question for a wide range of RBPs, with both offering important insights into RNA binding motifs and multivalency. Although often represented as sequence logos (Schneider and Stephens, 1990), it is important not to underestimate the importance of RNA structure as well as sequence in determining RBP binding. RNA is not a linear structure. As soon as it is transcribed the nascent RNA starts to adopt secondary and eventually tertiary structures. Indeed this co-transcriptional folding is thought to account in part for the limitations of early structural prediction algorithms that relied on finding the most thermodynamically stable secondary
structure for a sequence considered in its entirety; newer algorithms incorporate this important biological consideration with improved results (Lai et al., 2013; Meyer, 2017; Proctor and Meyer, 2013). As the ribosome travels along the coding sequence, it displaces structures (and likely other RBPs) on its path, leading to further dynamics and unravelling of structures. There is no such juggernaut travelling through the length of the 3’ UTR, by comparison, and so structures that are present here are likely to play a larger role in interactions with RBPs.

Early in vitro methods, such as RNAcompete (Ray et al., 2013) used synthetic sequences to determine k-mer sequence binding specificities, while more recent methods, such as RNAcompete-S (Cook et al., 2017), RNA Bind-N-Seq (Dominguez et al., 2018) and high throughput RNA-SELEX (systematic evolution of ligands by exponential enrichment) (Jolma et al., 2018) now all consider sequence and structural features and have defined binding motifs for nearing 100 RBPs. Furthermore, the importance of considering both sequence and structure become apparent when considering algorithms to predict RBP binding patterns. Those that use structure, however crudely, generate better predictions and out-compete those that do not (Heller et al., 2017; Maticzka et al., 2014; Pan et al., 2018; Stražar et al., 2016). Of course, the role of RNA structures is critical when considering RBPs with dsRBDs, such as the Staufen family, and this is explored further in Chapters 4 and 5.

However, the relevance of assessing the full spectrum of motifs bound by an RBP, even at different affinities, is becoming increasingly apparent from both in vitro and in vivo methods; RBPs largely do not have a single ‘motif’. As a simple starting point, considering primary and secondary motifs can be useful (Begg and Burge, 2019; Jolma et al., 2018). For example, for the RNA-binding fox-1 homolog (Rbfox) splicing factor, only approximately half of its in vivo binding sites have the primary “GCAUG” or “GCACG” motifs. In depth in vitro characterisation of Rbfox binding motifs identified six additional novel 5-mers, to which it bound with moderate affinities; these secondary motifs can account for more than a quarter of the remaining binding sites identified in vivo (Begg and Burge, 2019). While the remainder of sites may be experimental noise, they may also reflect low affinity binding sites that remain more difficult to ascertain reliably.
Returning to the multi-domain structure of RBPs, this concept of different motifs is particularly relevant. This architecture highlights a role for conformational dynamics, as discussed earlier (Gallagher and Ramos, 2018; Lunde et al., 2007; Mackereth and Sattler, 2012). Hence, while short \(k\)-mers are important isolated motifs, many RBPs have more complex binding modes with spaced bipartite motifs and structural contexts (Dominguez et al., 2018). Moreover, different domains within the same RBP may have different sequence specificities, as for ZBP1. Genome-wide studies have supported structural work that co-operative interactions between multiple protein domains and clustered short RNA motifs can enable high-affinity binding over stretches of RNA (Cereda et al., 2014; Lunde et al., 2007; Zhang et al., 2013). Indeed the pattern of regulatory binding of a number of RBPs, such as NOVA, PTBP1 and TDP-43 implicates a key role for multivalency in functional protein-RNA interactions (Tollervey et al., 2011; Ule et al., 2006; Xue et al., 2009).

In vivo methods for assessing protein-RNA interactions (König et al., 2012; Lee and Ule, 2018) are discussed in more detail in Chapters 3, 4 and 5. Their major strength, however, lies in the ability to map the positional binding of RBP on transcripts genome-wide at high spatial resolution. Beyond simply ascertaining binding motifs (Cereda et al., 2014; Licatalosi et al., 2008; Tollervey et al., 2011; Ule et al., 2003), the location information enables integration with orthogonal data to derive position-dependent principles relating the binding patterns and mechanisms to the functional outcome of the interactions (Attig et al., 2018; Chakrabarti et al., 2018; König et al., 2010; Tollervey et al., 2011; Van Nostrand et al., 2017a; Witten and Ule, 2011; Xue et al., 2009; Zarnack et al., 2013). Beyond these well-established approaches, more recent methods have sought to incorporate the spatial organisation of protein-RNA interactions within cells, and stand to complement insights from high resolution binding profiles at the nucleic acid scale, with those from sub-cellular, even membrane-less compartments (Benhalevy et al., 2018; Kaewsapsak et al., 2017; Padrón et al., 2019).
1.2.6 RNA binding proteins and human disease

Given their central role in coordinating complex protein-RNA interaction networks that regulate RNA metabolism and consequently RNA function, it is not surprising that RBPs have important roles in human disease and in human health. Mutations in both the RBP and also in its target RNA binding site can result in disease. Furthermore, both loss-of-function and gain-of-function mutations can give rise to disease. For example, a CGG trinucleotide repeat expansion in the 5′ UTR of the FMR1 gene results in the loss-of-function of the RBP fragile X mental retardation protein for which it encodes, resulting in fragile X syndrome (Conlon and Manley, 2017). Conversely a CUG trinucleotide repeat expansion in the 3′ UTR of the DMPK gene, and therefore its transcribed mRNA, results in the entrapment of RBPs and gain-of-function of the RBP CUG-binding protein 1, which plays a part in the development of myotonic dystrophy type 1 (Conlon and Manley, 2017).

Disease-causing mutations in RBPs often manifest as neurological diseases; this may be related to the more complex RNA processing necessary in the brain as touched upon earlier in discussing post-transcriptional regulation in neurons. RBPs play important roles in processes such as alternative splicing, 3′ end processing and RNA localisation, which are prevalent in the brain as a major source of the functional neuronal complexity. Amyotrophic lateral sclerosis (ALS), also known as motor neuron disease, is one of the most studied diseases in this context and has given us many important insights into the mechanisms of action of a number of RBPs. Although only approximately 10% of ALS cases show familial inheritance (Chen et al., 2013), to date 29 genes have been implicated in both familial and sporadic forms of the disease: many of these are in RBPs, such as TDP-43, FUS and Matrin3 (Chia et al., 2018). TDP-43 aggregates are a pathological hallmark of ALS (even in cases without TBP-43 mutations), and understanding the implications of these is needed to unpick the pathophysiology of the disease process. Are they pathogenic or reactive or protective? TDP-43 tends to bind long UG-rich stretches that are important for splicing regulation (Polymenidou et al., 2011; Tollervey et al., 2011).
 CHAPTER 1. INTRODUCTION

However, TDP-43 also has an intrinsically disordered region, which is important for weak, repeated homomeric interactions that can result in aggregation of TDP-42 (Harrison and Shorter, 2017). Mutations within these regions can affect its ability to phase separate and to form membraneless organelles (Conicella et al., 2019). Thus this may also be a relevant disease-causing pathway.

One important and topical example of how an understanding of RNA-protein interactions has led to major a medical advance is the disease spinal muscular atrophy (SMA) and the remarkable recently approved treatment nusinersin, which can generate significant improvements in disease symptoms (Bennett et al., 2019). SMA is a neurodegenerative disease that targets motor neurons, manifesting usually in the early months of life. Although there are different types of SMA, with different severities, it is not uncommon for patients not to survive through childhood. The disease is caused by a loss of function mutation in the SMN1 gene which, together with SMN2 in humans, encodes the survival of motor neuron (SMN) protein. SMN is not an RBP, but is critical for the assembly of Sm proteins and small nuclear RNAs into RNPs. SMN2 arose from a duplication event and differs from SMN1 by 5–13 nucleotides. Notably though, it only produces a minority of SMN owing to a translationally silent C to T single nucleotide polymorphism in exon 7 (Bennett et al., 2019; Conlon and Manley, 2017). However, this inactivates a splicing enhancer and instead introduces a splicing silencer that results in skipping of exon 7 which then produces a non-functional transcript that is degraded (Cartegni and Krainer, 2002; Conlon and Manley, 2017). This splicing silencer is bound by the RBPs hnRNP A1 and Sam68. Boosting levels of SMN in the context of mutations in SMN1 by preventing exon-skipping in SMN2 by blocking the action of the repressive RBPs and thus producing a functional transcript was therefore an attractive proposition. Screens of antisense oligonucleotides (ASOs) targeting this and other regulatory regions were evaluated with the most effective ASO, ultimately named nusinersin, taken forward: in mice it was found to include exon 7 in a dose-dependent manner and improve survival and prevent motor neuron and muscle pathology in disease models, with subsequent clinical trials not only establishing its role in the neurodegenerative therapeutic arsenal, but also leading the way for ASOs in other disease
1.2. PROTEIN-RNA INTERACTIONS

(Bennett et al., 2019).

Although I have focused on neurological diseases, RBPs also have mutations and altered expression levels and probably play important roles in a number of cancers (Cooper et al., 2009). However, even beyond disease causing mutations, it is becoming more widely appreciated that RBPs control cellular phenotype changes that occur during responses to acute or chronic injuries in many organs. For example, their post-transcriptional regulatory roles are important in the cardiovascular system in diseases such as atherosclerosis, or cardiac hypertrophy, or cardiac remodelling after myocardial infarction (de Bruin et al., 2017). In mouse models, knockdown or knockout of a number of RBPs can result in a range of cardiovascular developmental defects (de Bruin et al., 2017). Many developmental processes are reactivated after injury as part of the repair process. However, this response can also result in pathophysiological consequences, and the RBPs that coordinate splicing events in key regulatory genes in developmental and regenerative or repair processes are only now starting to be investigated (de Bruin et al., 2017). In the future there may even be therapeutic avenues by harnessing or modulating RBP actions at key time points in cardiovascular repair to reduce the likelihood of deleterious consequences. The scope of studying RBPs is gradually broadening across the spectrum of human diseases, ranging from metabolic disorders (Salem et al., 2019) to viral infections (Garcia-Moreno et al., 2019).

So, developing a detailed understanding of protein-RNA interactions at a molecular level has important implications for our understanding of human health and disease. This is necessary ultimately to develop treatments for many devastating neurological diseases that are currently without cure. The potential of this approach is exemplified by the recently approved ASO treatment for SMA. Moreover, modulating RBP function may also be a useful approach to tackling deleterious effects of appropriate repair responses across the spectrum of human organ systems. RBPs do not work in isolation but in complex regulatory networks and so studies in health and in disease, and in relating pathogenic molecular events to their phenotypic effects will be necessary to map these networks.
1.3 INTEGRATION OF GENOMIC APPROACHES

To study RNA-protein interactions a number of experimental methods have been developed, both genome-wide (König et al., 2012; Lee and Ule, 2018) and proteome-wide (Hentze et al., 2018) offering complementary RNA-centric and protein-centric viewpoints (Hentze et al., 2018; Ramanathan et al., 2019). Similarly, there are a range of DNA-centric and more protein-centric methods to study DNA-protein interactions (Klemm et al., 2019; Park, 2009). In this thesis I have primarily taken the nucleic acid viewpoint. I introduce the relevant genome-wide experimental methods in more detail within each chapter. Here, I briefly consider what it means to “integrate” data and why such an approach is fruitful.

An overwhelming amount of sequencing data has been generated, from a range of different flavours of functional genomic experiments\(^4\). From the perspective of trying to piece together an understanding of biological systems, each of these experiments is only a piece of the jigsaw puzzle. It is necessary to take multiple angles to study a biological question. Dissecting the example of the importance of protein-RNA interactions to neuronal function discussed in Section 1.2.3 provides a good example of the utility of integrating data sets. 3′ end sequencing methods can identify alternative 3′ isoforms; RNA sequencing can identify alternative splicing events; sequencing methods to measure RNA stability can assess the metabolism of those isoforms; RBP binding studies can determine sites of interaction on RNA molecules. Assimilating information from all these kinds of experiments are necessary to build up a molecular mechanistic picture of neuronal function. Moreover, it can be highly valuable to leverage the wealth of publicly available data already generated to support, complement and augment data produced by experiments for a given study.

\(^4\)A search for the term “rna seq” in The National Centre for Biotechnology Information’s Sequence Read Archive, only one repository for sequencing data, had 1,577,517 hits on 30th August 2019
1.4 OVERVIEW OF THE THESIS

In this thesis I present my studies into protein-nucleic acid interactions, considering methods to study them, mechanisms that guide them, and functions that are regulated by them.

In Chapter 2, I present our work into a type of interaction between eukaryotic DNA and a protein that, although exogenous, has revolutionised scientific research and will likely in the future have major therapeutic implications: Cas9. We explore the nature of CRISPR-Cas9 editing interactions at over 1,000 genomic targets and show that the outcome can be predicted by simple rules that mainly depend on the target sequence. We also investigate how the chromatin environment affects these outcomes.

Then, for the subsequent chapters, I focus on interactions between RNA and proteins. I start with evaluating and developing methods to study them. In Chapter 3, I focus on different variants of CLIP technologies and approaches to analyse the data they produce. I establish a set of standards that ensure robust assessment of data quality before drawing biological interpretations. I then study how the use and parameterisation of different bioinformatic tools can affect the results. In Chapter 4, I examine the hiCLIP method to study interactions between proteins and RNA structures. I identify three shortfalls in the data analysis and develop computational methods to overcome them, leading to an expanded, more robust data set to enable further study.

Finally, in Chapter 5, I turn to developing biological insight into the nature of interactions between RNA structures, notably duplexes, and the Staufen family of RNA binding proteins. In the first part, I show the utility of the improved data set from chapter 4 to contextualise the interactions within the diverse RNA structures present in cells and generate functional hypotheses for further investigation. In the second part, I analyse new experiments from rat brain tissue at different stages of post-partum development to study the roles of these RBPs in a more physiological context, in a highly relevant cell-type.
CHAPTER 1. INTRODUCTION

Throughout, backstage to the protein-nucleic acid interactions under study, but in a key supportive role, are integrative computational analyses: the interactions between different types of data, both newly generated and publicly available. In Chapter 2, I use ChIP-seq and DNase-seq data to support some of the findings from our DNA-seq experiments. In Chapter 3 I use RNA-seq and splicing microarray data after protein knock-down/out to evaluate CLIP data and its analysis. In Chapter 4, the methods I develop use data generated by computational predictions to augment and validate the hiCLIP data. And in Chapter 5, I integrate hiCLIP data with a wide range of functional genomic data sets that study expression, 3′-end processing, stability, structure and other protein-RNA interactions.
# Insights into Interactions between Eukaryotic DNA and CRISPR-Cas9

## Chapter 2

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Summary of the chapter</td>
<td>56</td>
</tr>
<tr>
<td>2.2</td>
<td>Introduction</td>
<td>56</td>
</tr>
<tr>
<td>2.2.1</td>
<td>What is genome editing?</td>
<td>56</td>
</tr>
<tr>
<td>2.2.2</td>
<td>A brief history of CRISPR-Cas9</td>
<td>57</td>
</tr>
<tr>
<td>2.2.3</td>
<td>DNA repair pathways</td>
<td>61</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Characterising on and off target effects</td>
<td>63</td>
</tr>
<tr>
<td>2.3</td>
<td>Methods</td>
<td>65</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Read alignment and indel identification</td>
<td>65</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Characterisation of target sites</td>
<td>67</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Analysis of the van Overbeek et al. (2016) data</td>
<td>69</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Assessing the effects of chromatin modulation on indel profiles</td>
<td>69</td>
</tr>
<tr>
<td>2.3.5</td>
<td>Analysis of the chromatin environment</td>
<td>70</td>
</tr>
<tr>
<td>2.4</td>
<td>Results</td>
<td>70</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Robust identification of indels</td>
<td>70</td>
</tr>
<tr>
<td>2.4.2</td>
<td>A large scale analysis of indel profiles</td>
<td>73</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Indels are reproducible and non-random</td>
<td>74</td>
</tr>
<tr>
<td>2.4.4</td>
<td>General patterns of indel profiles</td>
<td>75</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Precision of editing: what is it?</td>
<td>76</td>
</tr>
<tr>
<td>2.4.6</td>
<td>The role of sequence homology in guiding repair</td>
<td>78</td>
</tr>
<tr>
<td>2.4.7</td>
<td>DNA sequence determinants of editing precision</td>
<td>83</td>
</tr>
<tr>
<td>2.4.8</td>
<td>Sequence content of the precision core</td>
<td>84</td>
</tr>
<tr>
<td>2.4.9</td>
<td>Chromatin modulation affects editing efficiency</td>
<td>86</td>
</tr>
<tr>
<td>2.4.10</td>
<td>Chromatin states influence indel profiles</td>
<td>90</td>
</tr>
<tr>
<td>2.5</td>
<td>Discussion</td>
<td>91</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Precision of editing outcome</td>
<td>91</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Different routes to precision: complementary studies</td>
<td>93</td>
</tr>
<tr>
<td>2.5.3</td>
<td>The relationship between precision and indel type</td>
<td>94</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Nucleotide –4 plays a key role in defining site-specific indel profiles</td>
<td>95</td>
</tr>
<tr>
<td>2.5.5</td>
<td>The influence of the chromatin environment on site-specific editing outcomes</td>
<td>96</td>
</tr>
</tbody>
</table>
2.1 SUMMARY OF THE CHAPTER

In this chapter I present our work into studying the patterns of outcomes after CRISPR-Cas9 editing of genomic DNA. I start by discussing gene editing methods and how CRISPR-Cas systems fit into this technological toolbox. Then I introduce the interplay between the Cas9 enzyme and the cellular DNA repair machinery that ultimately results in editing. We examined indel profiles at over 1,000 sites and found that precision of gene editing, that is the recurrence of a specific indel, varies considerably between sites. We discovered that sequence homology plays an important role in determining the types of insertion or deletion, and that the indel profiles at precise target sites are robust to differences in chromatin or cell type. Moreover, in addition to charactering these profiles, through analysis of sequence composition, we discovered simple rules that can be used to predict the precision of editing outcome.

This work was a collaboration with Mr Tristan Henser-Brownhill and Dr Josep Monserrat from the Scaffidi Lab. The experiments were performed by Mr Tristan Henser-Brownhill, Dr Josep Monserrat and Dr Paola Scaffidi. I designed and performed all of the main computational analyses, except for the neural network modelling and permutation nucleotide importance analysis which was carried out by Mr Tristan Henser-Brownhill. The work presented here has been previously published and is adapted from Chakrabarti et al. (2019).

2.2 INTRODUCTION

2.2.1 What is genome editing?

Genome engineering has been pursued for decades: ultimately there is the tantalising prospect of therapy for genetic diseases. Although we stand on the cusp of applications in plant and human sciences, currently its mainstay is as a powerful tool for making fundamental discoveries in biology. Gene editing is the process of
altering the DNA sequence of a targeted gene by means of deleting, inserting or modifying nucleotides. These methods generally use targeted nucleases to generate double strand breaks (DSBs) prompting repair of the damaged locus. Although the endogenous DNA repair machinery has a remarkably high level of fidelity, it is not perfect and consequently can introduce errors, leading to editing the sequence of the target gene.

There are two processes that need to occur to create these targeted double strand breaks necessary for gene editing. First, there needs to be an enzyme that can cleave the DNA molecule and second, and more crucially, there needs to be a method of directing that enzyme to a specific target sequence reliably. Early genome engineering approaches revolved around zinc finger proteins (ZFs) and transcription activator-like effectors (TALEs) that have DNA binding domains that recognise 3 and 1 bp of DNA respectively. Modules of these can be assembled to target defined sequences (Zhang et al., 2011). Furthermore, they can be fused to an endonuclease to generate site-specific nucleases (Hsu et al., 2014). While technologies based on engineered ZF or TALE nucleases have been successfully employed for editing (Carroll, 2011; Miller et al., 2011), they are highly labour-intensive, requiring the synthesis of new protein conglomerates for each target sequence, with many limitations ranging from context-dependent effects and difficulties with their delivery into human cells using lentiviral vectors (Sander and Joung, 2014). The subsequent discovery, characterisation and translation of a bacterial defence mechanism against viral infections has revolutionised the field.

2.2.2 A brief history of CRISPR-Cas9

Clustered regularly interspersed palindromic repeat (CRISPR) sequences were first noted in *Escherichia coli* bacteria. “An unusual structure was found... Five highly homologous sequences of 29 nucleotides were arranged as directed repeats with 32 nucleotides as spacing.” (Ishino et al., 1987). These CRISPR arrays were found in approximately 50% of bacteria and 90% of archaea (Makarova et al., 2015). All the
CHAPTER 2. INSIGHTS INTO INTERACTIONS BETWEEN EUKARYOTIC DNA AND CRISPR-CAS9

Figure 2.1: Schematic of the prokaryotic genomic CRISPR locus. The cas gene operon with the tracrRNA and the CRISPR array. CRISPR, clustered regularly interspersed palindromic repeat; Cas, CRISPR associated; tracrRNA, trans-activating CRISPR RNA. From Doudna and Charpentier (2014). Reprinted with permission from AAAS.

Key components of the CRISPR system can be found in the genomic CRISPR locus around the CRISPR array (Figure 2.1) (Doudna and Charpentier, 2014).

During the expression stage of CRISPR immunity, the CRISPR array is transcribed and processed into multiple individual CRISPR RNAs (crRNAs). In the bacterial genome, adjacent to the array, is an operon coding for the CRISPR associated (Cas) proteins. The Cas proteins are the effectors of CRISPR activity: the crRNAs guide these proteins to target and cleave specific corresponding viral sequences. The Cas proteins are highly diverse across species, but can be grouped into four functional modules: adaptation (spacer acquisition); expression (crRNA processing); interference (crRNA-target binding and target cleavage) and ancillary (regulatory and other CRISPR-associated functions) (Makarova et al., 2015). The CRISPR-Cas systems can be categorised using the proteins that make up the four modules. The CRISPR-Cas9 Type II system (from Streptococcus pyogenes in the early studies) is particularly tractable for biotechnological purposes as only one Cas protein, Cas9, can perform all the interference processes.

The crRNA, however, is not sufficient on its own to bind to the Cas9. A small trans-encoded RNA upstream of the Cas operon, the trans-activating crRNA (tracrRNA) is responsible for directing the maturation of the crRNA by RNase III and Cas9 (Deltcheva et al., 2011). The tracrRNA:crRNA duplex guides the Cas9 enzyme to the target site, but can be engineered as a single guide RNA (sgRNA) that retains a 20 nt sequence at the 5’ end that determines the complementary DNA target and a double stranded structure at the 3’ end that binds to the Cas9 protein (Doudna and Charpentier, 2014; Jinek et al., 2012; Sander and Joung, 2014).

So, despite the complexity of the many bacterial CRISPR-Cas immune systems,
the corresponding biological toolkit consists of just two key components: an sgRNA and a Cas9 protein. The system can be programmed, and the target altered, simply by changing the sequence of the sgRNA. Generating a single strand of RNA is both relatively cheap and technically simple. Moreover, the same protein molecule, Cas9, can be used for any target, without further engineering. Soon after CRISPR-Cas9 was shown to cut DNA \textit{in vitro}, it was successfully induced in eukaryotic cells to enable gene editing \textit{in vivo} (Cho et al., 2013; Cong et al., 2013; Mali et al., 2013). It is the ability of an sgRNA to direct a Cas protein to cleave DNA nearly anywhere in the genome of most organisms tested to-date that has revolutionised gene editing.

2.2.2.1 \textit{The mechanism of action of Cas9: the role of sequence}

The targeting specificity of the CRISPR-Cas9 system is driven by the sgRNA sequence. This is a 20 nt stretch of single-stranded RNA that is complementary to the complementary strand of a DNA target site, termed the protospacer. Usually the Cas9 enzyme is in an inactive state, but the loading of the sgRNA results in a first conformational change that creates a central channel to permit DNA entry and binding (Jinek et al., 2014).

The search for target sites is accelerated by the use of a short protospacer adjacent motif (PAM) that is 3′ to the protospacer. In bacteria this conserved sequence allows for discrimination between self and non-self when cleaving viral RNA. In the case of Cas9, this is a 5′-NGG trinucleotide. First, the Cas9 enzyme directly binds to the PAM (Sternberg et al., 2014). Then the 5′ stretch of DNA is assessed for complementary to the sgRNA (Sternberg et al., 2014; Wang et al., 2016). A range of methods, both \textit{in vivo} assays, such as mismatched guide libraries, and \textit{in vitro} assays, have been used to determine how much of the 20 nt sgRNA determines specificity. The first PAM-proximal 8–12 nt form a seed region that is critical for recognition of target sequences (Cong et al., 2013; Hsu et al., 2013; Jinek et al., 2012), with subsequent work showing that bases further away from them PAM are less important for specificity (Hsu et al., 2014, 2013). Mechanistically, this corresponds to a model where after Cas9 has bound the PAM, the DNA helix is gradually unwound, assessed for complementarity to the sgRNA, and an RNA-DNA
heteroduplex created if so (Sternberg et al., 2014; Wang et al., 2016).

After the Cas9-sgRNA complex is bound to the PAM and the protospacer, there is a second conformational change in the Cas9 enzyme that leads to activation of the nuclease domains (Sternberg et al., 2014). Cas9 creates a double strand break through the activity of two endonuclease domains: HNH and RuvC (Jinek et al., 2012): the active sites in the The HNH domain cleaves the DNA strand complementary to the sgRNA, while the RuvC domain cleaves the other strand (Figure 2.2).

2.2.2.2 **The retaliation of the eukaryotic genome: the role of structure**

Large scale studies have identified sequence patterns that correlate with sgRNA activity and so have generated algorithms guide optimal sgRNA design (Doench et al., 2016; Wang et al., 2014). Despite considerable improvements in *in silico* predictions of sgRNA efficacy, there is still only moderate concordance between predicted and observed indel activity (Henser-Brownhill et al., 2017). Hence, additional parameters not included in these algorithms likely contribute to the overall outcome.

The CRISPR-Cas9 system has evolved to act against viral DNA in a prokaryotic cell in which there are relatively short sequences of self-DNA that do not require further packaging. As discussed in Chapter 1, Eukaryotic DNA, however, forms
higher order structured complexes. Nucleosomes can directly impede the binding of Cas9 and prevent it from cleaving DNA (Horlbeck et al., 2016b) and incorporating data about nucleosome positioning improves the design of sgRNA for CRISPR interference libraries (Horlbeck et al., 2016a).

However, nucleosomes are not static structures, but in fact highly dynamic. In addition to alterations in their composition through post-translational modifications of the histones, their conformation can change with DNA wrapping and unwrap-ping, termed ‘breathing’ (Zhou et al., 2019). Conformational dynamics can influence Cas9 action, with in vitro experiments showing that greater nucleosomal breathing can enhance its action by transiently exposing an otherwise protected target site (Isaac et al., 2016). Compositional dynamics, namely changes in the chromatin status, also likely play a role, but have only been implicated through in vitro studies (Horlbeck et al., 2016b) or correlative evidence (Uusi-Mäkelä et al., 2018). As yet there is little robust evidence that the chromatin status of a locus in vivo affects editing potential or outcome, but is investigated further here.

2.2.3 DNA repair pathways

Inducing a DSB is only half of the story. Editing occurs as a result of the tension between two opposing mechanisms: the Cas9 enzyme cleaving DNA at a target site to create a break; and the DNA repair machinery acting to repair the break. Both of these compete at a target site until the target sequence can no longer be recognised by CRISPR-Cas9, and the cycle broken. It is important to note that the mutations are generated not directly by the Cas9 enzyme, but rather from the imperfect action of the DSB repair pathways. There are two key pathways relevant for our work: canonical non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ), which is one of the alternative end-joining pathways. The other major repair pathway, homologous recombination (HR) is much slower, and in the context of gene editing usually leveraged to assist in the introduction of a synthetically generated sequence of choice into a target locus.
2.2.3.1 Canonical non-homologous end joining

Canonical NHEJ is the direct ligation of the free ends at the site of a DSB. It is the commonest repair pathway to be invoked, accounting for approximately 75% of repair events. It is also usually the fastest, taking place within 30 minutes in one study (Her and Bunting 2018). By comparison, in the same study, the slower HR pathway took several hours. DSBs induced by Cas9 are predominantly repaired by the NHEJ pathway (Bothmer et al., 2017).

Blunt ends, such as those usually thought to be generated by Cas9, are generally repaired without end-processing, and directly ligated to each other (Chang et al.,...
2.2. INTRODUCTION

The Ku70-Ku80 protein heterodimer binds the DNA ends (Figure 2.3A), protecting the free ends and preventing extensive resection (Chang et al., 2017; Mimitou and Symington, 2010). Where the Artemis-DNA-dependent protein kinase complex does resect these incompatible ends (Figure 2.3B, C), typically this is by fewer than 20 nucleotides (Chang et al., 2017). Ku facilitates the binding of the XRCC4-DNA ligase IV complex, which in turn can bridge the two free ends and anneal them together to repair the DSB (Figure 2.3).

2.2.3.2 Microhomology-mediated end joining

MMEJ is sometimes considered as a back-up pathway when NHEJ has failed (McVey and Lee, 2008). If the Ku70-Ku80 heterodimer cannot bind, or is removed by an unknown mechanism, then the ends are unprotected and the nucleases Mre11 and CtIP can result in the 5′ to 3′ resection of one strand. Microhomologies, i.e. base complementarity within the two free single strand ends, can then be annealed together according to Watson-Crick base pairing to link the two overhanging ends. In mammalian cells there can be as little as 1 nt of homology (Sfeir and Symington, 2015) although usually a requirement for homologies of between 2 and 20 nt are described (Chang et al., 2017). After this, the flanking resected single-strand regions undergo fill-in synthesis by polymerase θ (Chang et al., 2017; Sfeir and Symington, 2015) with usually DNA ligase 3 finally sealing the ends (Chang et al., 2017; Sfeir and Symington, 2015). Microhomologies have been shown to play a role in CRISPR-Cas9 editing outcome (Bae et al., 2014) and from inhibition studies, we know that both NHEJ and MMEJ pathways contribute to repair and hence the editing profiles at a given target site (Brinkman et al., 2018; van Overbeek et al., 2016).

2.2.4 Characterising on and off target effects

There are two sides to the use of Cas9 as a gene editing tool: the enzyme needs to maximise efficiency of cleavage at on-target sites, while minimising effects at partially homologous off-target sites. Much research has been focused on charac-
Chapter 2. Insights into Interactions between Eukaryotic DNA and CRISPR-Cas9

terising and limiting off target effects, that is the unwanted editing of other loci containing sequences that are similar to the target site (Kim et al., 2019a). Many methods have been developed to assess the occurrence of off target effects including bioinformatic predictions, in vitro assessments using genomic DNA templates and in vivo sequencing methods to identify off target editing directly (Tsai et al., 2015).

However, large-scale characterisations of on-target editing profiles across the genome in vivo have been limited to date owing to two barriers: first, the need to transfect cells with multiple sgRNAs and second, the need to amplify and deeply sequence each target locus. Prior to our work, the largest survey of Cas9 editing was of 96 different sgRNA sequences (van Overbeek et al., 2016). Studying even this limited number of targets advanced our understanding of editing outcomes. Previously, as NHEJ was considered the predominant repair pathway and usually results in heterogeneous repair outcomes, editing resulting after CRISPR-Cas9 action were thought to be random. However, the indel patterns described by van Overbeek et al. (2016) showed that individual targets in fact had reproducible outcomes. By identifying sets of target loci across the genome that could be targeted by the same sgRNA, and assessing the indel profiles generated in all these regions they showed that a given sgRNA induced patterns of indels with distinct preferences for class (insertion or deletion) and size of indels at all of the target loci. This suggests that there is a deterministic nature to CRISPR-Cas9 induced DSB repair, however the factors involved in defining these non-random patterns remain unclear. Here, we performed a large-scale genomic characterisation of indel patterns examining over one thousand genomic sites in a human cell line, with the aim of understanding how genetic and epigenetic factors guide repair and consequently editing outcomes after CRISPR-Cas9 cleavage. This has enabled us to derive simple rules that guide the target-specific precision of CRISPR-mediated gene editing with implications for the use of this technology as both an experimental and potentially a therapeutic tool.
2.3. METHODS

The methods are taken and adapted from (Chakrabarti et al., 2019). The experimental set-up is summarised in Figure 2.4. There were 3 experiments with pools of 100 sgRNA, and 3 experiments, with 2 replicates each, with pools of 450 sgRNA. Here, I focus on the methods for the computational analysis I performed.

2.3.1 Read alignment and indel identification

FastQC (Andrews, 2010) was used to check the quality of the sequenced reads and ensure there was no adapter contamination. Reads were aligned using BBMap (v. 36.59) (Bushnell, 2015) to the GRCh37 genome assembly.

**INDEL IDENTIFICATION FOR THE LARGE-SCALE EXPERIMENTS**  
In order to identify the reads that contained indels robustly we adopted a two-stage alignment strategy. In the first phase we aligned the reads to the genome disallowing any reads that contained indels using the BBMap settings: `maxindel=0 strictmaxindel=t local=t`. We discarded reads that aligned in a proper pair in this phase and took the remainder forward. In the second phase we aligned the remaining reads to the genome, this time setting a soft threshold allowing indels up to 2000 bp using the settings: `maxindel=2000 local=t`. Duplicates were marked using Picard (v. 2.1.0). Reads that were marked as duplicates, or that had a mapping quality score of less than 38 were filtered using samtools (v. 1.2) and sambamba (v. 0.6.0). This two-phase approach was necessary to delineate, for a
given target amplicon, between reads from cells uninfected with the corresponding sgRNA and reads from cells with successful transfections, on account of the pooling of sgRNAs. For a given amplicon, only a small proportion of the total number of cells would have been transfected with the sgRNA targeting the site contained within it. We know that aligned reads containing indels arise from appropriately transfected cells. However, our approach forces the aligner to choose an alignment with no indels over one with indels for the multiple possibilities for a given read. With this approach we can improve our confidence that the reads with indels are not background noise or alignment errors. Because of the experimental approach, the sensitivity of our method is inherently limited, and it is likely that indels occurring at low frequency are not detected. Furthermore, kinetically slow repair events may be underrepresented in our dataset. Nevertheless, the observation that most targets are identified as imprecise or middle indicates that there is no significant bias towards most frequent indels.

Indel identification was performed in R (v. 3.4.4) using custom scripts. The location and size of indels in reads were identified from the CIGAR string. Indels were only considered valid if they occurred within 5 nucleotides of the Cas9 cleavage site (defined as 6 nucleotides upstream of the end of the guide RNA including the PAM sequence as in van Overbeek et al. (2016)). Any indels that could also be detected in the control HepG2 sample were removed as probable somatic mutations in this cancer cell line. To ensure robust estimate of indel frequencies, we filtered out target sites that had a low overall indel count (indels identified in fewer than 10 reads in total across all samples and replicates, where present).

ASSESSMENT OF THE INDEL IDENTIFICATION APPROACH   To assess possible confounding effects from sequencing errors, particularly given the depth of sequencing, we performed two complementary analyses. First, we assessed the number of indels detected at each target site (within 5 nucleotides of the Cas9 cleavage site) in the wild-type sample without Cas9 induction and sgRNA transduction (without filtering for probable HepG2 somatic mutations). Second, we leveraged the fact that all targeted regions in the whole library were pulled down and sequenced
to a similar depth in all experiments, irrespective of whether they were targeted in that particular pool or not. Therefore we compared the number of indels in both replicates from the 450 pool experiments at each target site in the experiment where the corresponding sgRNA was in the transfected pool, with the mean of the number of indels in both replicates from the two other 450 pool experiments where the corresponding sgRNA was not. This provided an estimate of the occurrence of sequencing errors in our experimental setup within 5 nucleotides of the Cas9 cleavage site.

**InDEL IDENTIFICATION FOR THE SMALL-SCALE VALIDATION EXPERIMENTS** Paired end reads were merged using BBMerge (v. 36.59) (Bushnell et al., 2017) using the settings k=150 qtrim=t trimq=10 adapter=AGATCG-GAAGAGC. Merged reads were aligned using the BBMap settings maxindel=2000 local=t. After alignment, duplicates were marked using Picard (v. 2.1.0). Reads that were marked as duplicates, or that had a mapping quality score of less than 38 were filtered as before using samtools (v. 1.2) and sambamba (v. 0.6.0). The R package CrispRVariants (Lindsay et al., 2016) was used to identify indels.

**2.3.2 Characterisation of target sites**

Throughout, we used all detected indels from both 100 and 450 pools to characterise the targets, except when assessing for efficiency where indels from the 450 pools only were used to ensure an unbiased analysis of each target site as explained above.

**Frameshifts and indel size** Indels were assessed for their frameshift potential by the divisibility of their size by 3. To identify patterns in the indel size profiles at target sites, we calculated the frequency of each size of indel (considered in bins of insertions greater than 1 nucleotide, insertions of 1, and deletions of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and greater than 10). We performed unsupervised hierarchical clustering using the Ward D2 method to categorise groups of target sites based on
their indel size profiles.

**PRECISION** We also categorised target sites by calculating the frequency of each distinct indel at each target site. The most frequent indel was termed the commonest; ties were broken by prioritising insertions over deletions, and then by longest deletion. The precision of indel generation at a target site was defined based on the frequency of the commonest indel: imprecise \( \leq 0.25 \), \( 0.25 < \text{middle} \leq 0.5 \), precise > 0.5.

**SEQUENCE HOMOLOGY** The presence of microhomology of \( n \) nucleotides was assessed in the deletions. The 5' \( n \) nucleotides of the deleted sequence were compared with the first \( n \) nucleotides downstream of the 3' join. Likewise, the 3' \( n \) nucleotides of the deleted sequence were compared with the last \( n \) nucleotides upstream of the 5' join. If there was a match, this was considered as microhomology. For each deletion sequence, values of \( n \) ranging from 1 to 50 (or the length of the deletion, whichever was shortest) were used. The largest matching \( n \) was considered the size of the microhomology.

Insertion homology was assessed by extracting the inserted nucleotide from the read sequence using the CIGAR string. This was compared with the nucleotide in the –4 position of the protospacer to assess for matches. When assessing the commonest insertion, we only considered target sites that had 5 or more insertions. Where the inserted nucleotide either creates, or lies within a short repetitive stretch; e.g. “A” inserted adjacent to “A” creating “AA”, or “T” inserted adjacent to/within “TT” creating “TTT”; it is not possible to identify precisely which of these nucleotides is the inserted position. The aligner arbitrarily assigns the first position to the inserted nucleotide.
2.3.3 Analysis of the van Overbeek et al. (2016) data

For the ‘spacer’ target sites from van Overbeek et al. (2016), aligned BAM files were obtained from the Sequence Read Archive for all time points after sgRNA nucleofection in HCT116, HepG2 and K562 cell lines. Indel identification was performed in R (v. 3.4.4) using custom scripts similar to before. Data presented here are from the longest, 48 hour time point. The location and size of indels in reads were identified from the CIGAR string. Indels were only considered valid if they occurred within 5 nucleotides of the Cas9 cleavage site (defined as 6 nucleotides upstream of the end of the guide RNA including the PAM sequence). For a given time point and cell type, indels that occurred with < 1% frequency were filtered, as were sites that had < 10% editing efficiency. Downstream analyses were performed as detailed in section 2.3.2.

2.3.4 Assessing the effects of chromatin modulation on indel profiles

Mutation efficiency was assessed using the mutationEfficiency function from CrispRVariants (Lindsay et al., 2016), considering single nucleotide variants as non-variants. To compare the counts of indels across the different conditions, in order to assess the contribution of each indel to the changes in efficiency, the raw counts for each indel in each condition were normalised to the library size for that condition. Indels that constituted less than 1% of the library size in any condition were filtered out.

To assess the effects of chromatin modulation on the indel profile of target, over and above the effects on efficiency, the raw counts were normalised differently by dividing by a size factor (the total number of indels detected in a condition). This permitted comparisons of the relative contribution of each indel to the overall indel profile across the different conditions. After normalisation, only the most frequent 10 indels in the untreated condition were used.
2.3.5 Analysis of the chromatin environment

DNase I hypersensitive sites sequencing (DNase-seq) and H3K9ac and H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) fold-enrichment data for in HepG2 cells were obtained pre-processed from Roadmap Epigenomics Consortium et al. (2015). I calculated the mean fold-enrichment signal in a 500 bp window centered on the cleavage site of the six validation targets.

Coverage files were obtained for DNase-seq, H3K9ac and H3K27ac ChIP-seq for HEK293 and K562 cell lines aligned to GRCh38 from Cistrome DB 43073, 45020, 45021, 45046, 55731, 58997 and GEO GSM1635901–6. 500 bp windows centered on the cleavage sites for the ‘spacer’ target sites from van Overbeek et al. (2016) were created and converted from GRCh37 to GRCh38 using the UCSC liftOver tool. The signal in each window was extracted using Deeptools and the mean calculated. These data were processed by Dr Anna Poetsch.

2.4 RESULTS

2.4.1 Robust identification of indels

In order to overcome the technical impracticalities of individually transfecting over 1,000 sgRNAs, a pooled strategy was employed to be able to target multiple sites within the population of cells. In order to increase the fraction of edited cells, a high multiplicity of infection of 10 was used. The major disadvantage of this pooled strategy was that most cells would not contain an sgRNA targeting a specific region. In part to compensate for this, sequencing of the targeted regions was performed to an ultra-high depth with read coverage of 6000 to 8000 over the amplicon around the target site (Figure 2.5A). This low signal to background ratio is apparent in the very high percentage of reads for each experiment that do not pass the uninfected or unedited threshold (Figure 2.5B).

With this depth of sequencing, however, there is the risk of conflating sequenc-
### 2.4. RESULTS

Figure 2.5: Sequencing library metrics

(A) Raw read coverage for each experiment of the region around the cleavage site that was selectively isolated by target capture.

(B) Alignment metrics for each experiment showing the number of reads at each stage of processing. From Chakrabarti et al. (2019).

...ing or alignment errors with valid editing events. To combat this and robustly identify the reads that contained indels, I adopted a two-stage alignment strategy. This was necessary to delineate, for a given amplicon, between reads from cells uninfected with the corresponding sgRNA and reads from cells with successful transfections.

Amplicon reads containing indels could arise from three sources:

1. Editing occurring at sites in cells in which the corresponding sgRNA was successfully transfected
2. Somatic mutations in the HepG2 cancer cell line
3. Sequencing or alignment errors

The challenge here was to disambiguate 1 from 2 and 3. To exclude confounding somatic mutations, a wild-type HepG2 sample without any Cas9 induction was used to create a catalogue of possible indels that were somatic mutations in this cancer cell line. Any indels detected in the experiments that were also present in this control were filtered out.

To reduce the possibility of spurious alignment errors, I forced the aligner to choose an alignment with no indels over one with indels for a given read by adopting a two-stage alignment strategy. Reducing such errors was also crucial as the identification of the location, size and type of indel was from the alignment. The side-effect of this approach was a loss of the ability to calculate exact editing
efficiency, as it was not possible to disambiguate from the unedited reads at a target site how many were from uninfected cells, and how many were from cells that were infected, but remained unedited. Given the high depth of sequencing, it was a concern that sequencing errors may be erroneously called as indel events. The quality control metrics for our experiments show a the mean Phred quality score was at least greater than 30 across the reads\(^1\). However, incorrect base calls usually result in an incorrect nucleotide being assigned, rather than one being missed or inserted. As only indels are considered editing events, this further reduces the probability of false positives.

Furthermore, there were two ways of demonstrating that the experimental and computational methods adequately control for false positives. First, the control wild-type HepG2 showed the rate of indels detected within a ±5 nt of the cleavage site (Figure 2.6B). For the vast majority of sites, no indels were detected. The tiny minority in which indels were detected reflect a combination of somatic mutations and false positive.

Second, and with greater internal validity, I leveraged the pooled experimental setup. The target capture probes were designed for all targets across the whole library. So, DNA from all sites (both targeted and untargeted) were pulled down

\(^1\)A Phred score of 30 equates to a 1 in 1000 probability of an incorrect base call.
and underwent ultra-deep sequencing to a similar depth for each experiment. For consistency, I focused only on the 450 pool experiments. For each target site, I compared the number of indels in both replicates in the experiment where the corresponding sgRNA was in the transfected pool, with the mean of the number of indels in both replicates of the two other experiments where the corresponding sgRNA was not (Figure 2.6A). The number of indels detected in control samples was a minor fraction of that detected in the targeted samples, and also falls below the 10 read threshold for the vast majority of sites. Given that we required a threshold of 10 indels to be detected at a site for it to be included in the downstream analysis, the number of indels detected exceeds background, hence I conclude that our approach robustly controls for false positive arising from sequencing errors.

Finally, further validation of our findings using an independent dataset (van Overbeek et al., 2016), in which individual rather than pooled sgRNA were used in addition to an alternative alignment method and tool lends additional support to our strategy.

2.4.2 A large scale analysis of indel profiles

We selected 1,491 target sites across the genome and used sgRNA from a previously generated and well-characterised lentiviral library that has been shown to have high activity (Henser-Brownhill et al., 2017). We selected 450 genes to target, with at least three sites per gene, distributed across the length of the gene. Indels were detected at 1,248 of the 1,491 target sites, ranging from 1 to 188 per target, with a median number of 32 at a target site (Figure 2.6A). On account of the pooled experimental strategy there was a limited sensitivity to detect all induced indels, particularly affecting sites with lower editing efficiency, this number was likely an underestimate. Nevertheless, there are sufficient indels to identify general patterns. This is supported by comparing between indel profiles generated in the large-scale analysis with those from six validation experiments where cells were transfected with individual sgRNAs (Figure 2.7).
2.4.3 Indels are reproducible and non-random

Editing was reproducible across two biological replicates, consistent with previous smaller-scale studies showing that editing outcome were non-random (Brinkman et al., 2014; van Overbeek et al., 2016). This was evident both at the study-wide level where comparing the frequencies of a given indel at a given target in both replicates (Spearman’s coefficient, 0.75) (Figure 2.8A) and also at the individual target level (Figure 2.8B). Even in the validation experiments, where the indel profiles could be assessed to a much finer resolution, there was remarkable reproducibility between two independent experiments (data not shown, Chakrabarti et al. (2019)).

Figure 2.8: Comparison of large-scale and validation experiments
(A) The frequency at which each detected indel occurs at each target site in two biological replicates.
(B) Indel profiles for two biological replicates for three example target sites. Indel nomenclature is [start coordinate relative to cleavage site]:[size][insertion or deletion]. Counts are normalised to the total library size for each experiment. Numbers in grey indicate indel frequency. From Chakrabarti et al. (2019).
2.4. RESULTS

Figure 2.9: The sizes and frame-shifting effects of indels
(A) The size distribution of the commonest indel at each target.
(B) The percentage of mutations resulting in a frameshift mutation at each target. The inset pie chart shows the proportion of targets for which the commonest indel induced a frameshift. From Chakrabarti et al. (2019).

2.4.4 General patterns of indel profiles

With this compendium of indel profiles at a large number of sites, we were able to make several general observations on the patterns of CRISPR-mediated indels. First, single nucleotide indels were the most frequent type of indel, with 44% of targets having a 1 nt insertion as the commonest indel and 26% having a 1 nt deletion. However, there were sites showing a preference for longer deletions of up to 41 nt as well (Figure 2.9A). Second, frameshift mutations were over-represented: on average, 80.1% of indels at a given target site resulted in a frameshift, much higher than the theoretical 66% assuming a random outcome (Figure 2.9B). Overall 81% of all detected indels results in a frameshift. This is consistent with the bias for single nucleotide indels. However, it is worth noting that for three sites, there was a predominance of in-frame indels (in over 70% of indels). This suggests that while the probability of inducing a loss of function at the protein level is relatively high, there are some sites that may be more resistant and so in order to knockout some genes it may be beneficial to target multiple sites. Third, unsupervised hierarchical clustering of the the distribution of sizes of indels at target sites, categorised sites into four main indel preference groups: i) insertions, ii) short deletions, iii) long deletions, and iv) mixed or no preference (Figure 2.10). Fourth, sgRNA activity at a site was highly variable, ranging from 0 to 188 indels per site and did not correlate with preference for a particular indel size. (As discussed earlier, it was not possible to assess editing efficiency on account of the pooling strategy.) Activity is likely
intrinsically variable as sgRNA activity did not correlate with the abundance of sgRNA in the pools, consistent with previous studies attempting to infer sgRNA activity from the ability to induce an expected phenotype (Doench et al., 2014; Wang et al., 2014).

2.4.5 Precision of editing: what is it?

Building on previous work (van Overbeek et al., 2016) and our observations concerning the non-random nature of editing outcome, we decided to explore the recurrence of a specific commonest indel, across sites. We termed this precision. To examine the precision of editing, we calculated the frequency of each distinct indel, defined based on both its coordinates and, for insertions, base composition. Ranking target sites based on the frequency of the commonest indel observed at that site revealed a striking pattern (Figure 2.11). Overall, there was a wide range of editing precision: ranging from some sites showing one dominant indel (commonest indel frequency up to 94%) and others displaying multiple infrequent indels (up to 79 distinct indels, with frequencies < 5%) (Figures 2.11 and 2.12). Overall, for approximately one-fifth of our targets there was at least a 50% probability of inducing a specific indel, but for the majority the outcome was more unpredictable. On average, the commonest indel frequency for a given site was 34.1% and the median number of distinct indels...
2.4. RESULTS

Figure 2.11: Site-specific precision of editing
(A) The heatmap visualises the frequency of each indel at each target site. In red is the commonest indel for that site, while in shades of blue are the next commonest ranked from 2–19. Indels ranking higher than 20 are in grey. The bar plot about shows the frequency of distinct indels at each target site below.
(B) Indel profiles at four target sites. The two on the left are imprecise targets, corresponding to the left of the heatmap in (A) and the two on the right are precise targets, corresponding to the right. Indels are ordered by start coordinate relative to the cleavage site (arrowhead), with insertions given priority over deletions. The inset number is the total number of indels at that site. From Chakrabarti et al. (2019).

We categorised target sites into three groups based on the frequency of the commonest indel: imprecise (0 to 0.25), middle (0.25 to 0.5), and precise (> 0.5) (Figure 2.13A). Each group contained broadly similar number of sites, and enabled a more detailed exploration of the relationship between precision and target site editing characteristics. First, using total number of indels observed at a target as a proxy for editing efficiency, we observed a positive correlation between precision and efficiency (Figure 2.13B). Precise targets showed on average twice as many indels as imprecise targets. This observation was not confounded by either differences in sgRNA abundance or sequencing depth (data not shown, Chakrabarti et al. (2019)). Second, we observed that indels at imprecise targets were largely deletions with only 20% of indels being insertions, but insertions were more frequent with increasing
precision (Figure 2.13C). Notably, precise targets could be divided into two very distinct subsets: insertion-preferred (68.4%) and deletion-preferred (31.6%). Third, when considering the size of the indels, there was a significant negative correlation between absolute size and precision (Figure 2.13D). Imprecise targets showed a range of indel sizes, with deletions up to 2 kilobases in length. However, precise targets had a very strong preference for single nucleotide indels, with 71.5% having a one nucleotide insertion or deletion (Figure 2.13D, E).

2.4.6 The role of sequence homology in guiding repair

As discussed earlier, it has previously been shown that DNA repair after CRISPR-Cas9 editing occurs through repair pathways that either do (MMEJ), or do not (NHEJ), depend on sequence homology in the absence of a template. We performed a pathway-agnostic assessment of the role of sequence homology in driving the indel profiles, by examining the indel boundaries for any homology. Although it was not possible to relate this back to specific pathway mechanisms, it nevertheless provided useful insight into the relationship between sequence and precision, both for deletions and insertions.
Figure 2.13: Relationships between editing precision and features of indels
(A) Distribution of commonest indel frequencies. The background indicates three groups of sites based on their precision of editing. Inset numbers indicated the number of target sites in the group.
(B–D) Relationship between precision group and indel count (B), type of indel (C) and indel size (D). I, imprecise; M, middle; P, precise. Statistical comparisons were done using the Kruskall-Wallis test followed by Dunn’s test with Benjamini-Hochberg correction for multiple comparisons.
(E) Complementary to (D), the relationship between indel size and commonest indel frequency (i.e. precision). From Chakrabarti et al. (2019).
Figure 2.14: Microhomology-associated deletions

(A) A representative target site is shown with all the different types of deletion detected. In the grey panel is the reference sequence, with the PAM emboled in blue and the cleavage site indicated by the red line. Below each line is a detected type of deletion: in the dashed box is the microhomology in the deleted sequence and emboled in red is the corresponding microhomology in the unedited sequence.

(B) The percentage of microhomology-associated deletions at each target site. The inset pie chart shows the proportion of all detected deletions that show microhomology.

(C) The percentage of deletions with microhomology of a given size. The grey bar indicates the expected frequency for each k-mer size. Statistical comparisons were done using the χ² test.

(D) The percentage of microhomology-associated deletions at individual sites grouped by precision. I, imprecise; M, middle; P, precise; MH, microhomology-associated. Statistical comparisons were done using the Kruskall-Wallis test followed by Dunn’s test with Benjamini-Hochberg correction for multiple comparisons. From Chakrabarti et al. (2019).
2.4. RESULTS

Figure 2.15: Microhomology-associated insertions
(A) The frequency of the commonest insertion at a given target site. The inset number indicates the number sites; those with fewer than 5 insertions were excluded to avoid a low-count bias.
(B) Representative targets site is shown with detected insertions. In the grey panel is the reference sequence, with the PAM emboldened in blue and the cleavage site indicated by the red line. Below the edited sequence is show with the insertion homology emboldened in red. From Chakrabarti et al. (2019).

An example of the ARID1B.7 target shows the nature of the microhomology-mediated deletions at a target site (Figure 2.14A). For deletions, we observed microhomologies ranging from 1 to 18 nucleotides. Overall, 73.3% of all deletions across all sites showed evidence of microhomology mediated repair. This corresponded to 74.3% of all deletions at a given site, on average (Figure 2.14B). The MMEJ pathway usually uses longer microhomologies ranging from 5 to 25 nucleotides (Deriano and Roth, 2013), but we observed that even short microhomologies of 1 to 4 nucleotides were enriched over the expected frequencies (Figure 2.14C). Furthermore, a significantly higher percentage of microhomology-mediated deletions were observed in the precise and middle targets compared with the imprecise ones (Figure 2.14D). However, regardless of precision, 80% of targets had a microhomologous deletion as their commonest.

We also observed sequence homology in insertions, although this had not previously been implicated in the repair mechanisms. For many sites, the inserted nucleotide appeared non-random, with a significant skew in the number of targets with a highly frequent commonest inserted nucleotide (Figure 2.15A). This recurrently inserted nucleotide was often homologous to the nucleotide in position –4 upstream of the PAM sequence; four representative examples are shown in Figure 2.15B. This is the nucleotide immediately 5′ to the cleavage site. We termed this
feature ‘insertion homology’ and observed it for 82% of the commonest insertions at each target. As for deletions, there was a significant positive correlation between the prevalence of insertion homology and precision: 96% of insertions showed homology for precise targets, contrasting with 57% for imprecise (Figure 2.16A). By comparing the rates of insertion homology for all insertions to those for the commonest insertions, one can see that even for imprecise targets, homologous insertions are often the commonest ones (Figure 2.16B). Finally, precise targets showed a strong preference for inserted “A”s and “T”s (Figure 2.16C), while there was a more even distribution of nucleotides for imprecise targets. Altogether, this suggests that template-mediated insertions are an important factor determining the site-specific indel profiles.

Overall, homology mediated end-joining has a major impact on DNA repair after Cas9 cleavage both for insertions and deletions, and correlates with the precision of editing.
Figure 2.17: A neural network identifies protospacer nucleotide positions important for editing outcome

(A) Sequence logo showing the distribution of nucleotides at each position of the sgRNAs from our experiments and from van Overbeek et al. (2016).

(B) The contribution of the indicated protospacer nucleotide to editing precision. The effect of randomising the nucleotide on the prediction of precision is shown as the the reduction of the model’s accuracy R^2. (Panel produced by Mr Henser-Brownhill.) From Chakrabarti et al. (2019).

2.4.7 DNA sequence determinants of editing precision

The strong relationship between insertion homology and precision led us to examine how the base composition of the target site related to precision. To try to unravel a possible relationship between the sequence content of the target site and precision, we used a machine learning approach. We trained a neural network to predict precision, i.e. the frequency of the commonest indel, with the 20 nucleotides of the protospacer sequence as the input. We also calculated precision for the independent van Overbeek et al. (2016) data set that characterised indel profiles for 96 target sites. This dataset had sgRNA that were selected using a different method to ours and showed overall different nucleotide composition (Figure 2.17A). Regardless of this, the trained network was able to predict precision equally well, suggesting firstly, that the sequence did relate to precision and secondly, that our model had learnt generalisable features.

The predictive power of the network was only moderate (coefficient of determination, R^2 = 0.49), however this was sufficient for our aim of identifying positions in the protospacer critical for determining precision. We performed a permutation ‘nucleotide’ importance analysis to do so. Our approach was to randomise each nucleotide in the protospacer in turn (i.e. replacing it with an “N”) and then assess the impact of this randomisation on the ability of the model to predict precision.
Figure 2.18: The nucleotide content of the precision core

Sequence logos for the precision core for the different precision groups. Precise targets are divided based on their preference for insertions or deletions. The most important –4 nucleotide position is indicated by the yellow box. From Chakrabarti et al. (2019).

(measured by the effect on the correlation between estimated and observed commonest frequencies). From this analysis, the nucleotide at position –4 upstream of the PAM sequence had by far the strongest impact on precision as an isolated position, reducing the model’s performance by 78% ± 9% (Figure 2.17B). The nucleotides at positions –2, –3 and –5 also had an effect, albeit weaker, reducing $R^2$ by 29% ± 9%, 15% ± 5% and 50% ± 13%. Furthermore, simultaneously randomising all 4 positions reduced $R^2$ by 98% ± 2%, and essentially abolished the predictive power of the model, indicating that these 4 positions are critical for defining the precision of a target site. We referred to them as the ‘precision core’. And within the precision core the position immediately upstream of the cleavage site, at position –4, is the most important.

2.4.8 Sequence content of the precision core

Having identified the precision core, we sought to understand how the sequence composition of the core related to precision by examining the sequence content of the core for each precision group (Figure 2.18). This confirmed that the nucleotide at position –4 was the most discriminating, followed by position –5 which also commonly had a “C” at precise targets.

We then focused on the nucleotide at position –4: different nucleotides showed particular associations with indel types (insertions versus deletions) and precision groups (Figure 2.19A, B, C). The overwhelming majority of target sites with an
Figure 2.19: The nucleotide at position –4 and editing outcome

(A) The proportion of target sizes with the indicated nucleotide at position –4 in groups of precision and commonest indel preference for insertion or deletions.

(B) The percentage of target sites grouped based on the nucleotide at position –4 showing a commonest indel preference for insertions of deletions.

(C) The precision of targets categorised by the nucleotide at position –4, in HepG2 (our data, highlighted in yellow) and three other cell lines (van Overbeek et al., 2016).

(D) As in (B) right panel, but for van Overbeek et al. (2016). From Chakrabarti et al. (2019).

“A” or a “T” in this position resulted in an inserted nucleotide upon repair of the double-strand break (77% and 91% of targets respectively). A “G” in this position however, mostly resulted in a deletion (79% of targets). Furthermore, for “A” and “T”, the target sites predominantly fell into the precise or middle groups, with median commonest indel frequencies of 0.42 and 0.56 respectively, whereas for “G” they were largely imprecise, with a median commonest indel frequency of 0.21.

To validate our results, we again turned to the van Overbeek et al. (2016) data and examined how precision and indel type related to the nucleotide at position –4 (Figure 2.19C, D). Broadly, similar distributions were observed for the 96 different sites across three different cell lines to ours, particularly so for the HCT116 and HEK293 cells (from solid organs, similar to our HepG2 cells).

Considering the nucleotide at position –5 in addition to –4, offered a further
refinement, consistent with the sequence content results (Figure 2.20). A “CA” or an “AT” increased the frequency of a precise target to 0.53 and 0.65 respectively, and these were mostly insertions. A “CC”, however, resulted in mostly deletions (74.6% of targets) at targets that were relatively precise (mean commonest indel frequency of 0.39).

Overall, we can predict the editing outcomes by assessing the nucleotides comprising the precision core, and the nucleotide at position –4 is sufficient to predict whether editing at a target site will result in insertions or deletions.

2.4.9 Chromatin modulation affects editing efficiency

Despite the strong relationship between DNA sequence and editing outcome that we observed, two findings suggested that in addition there were other factors at work. First, the neural network model, although able to predict sufficiently well enough based on sequence for us to determine the key position, was overall a mediocre predictive tool and unable to recapitulate fully the observed precision rates. Secondly, the number of indels and the indel profiles varied across groups. Given previous studies highlighting the importance of nucleosome positioning in editing efficiency (Horlbeck et al., 2016a,b), we explored the contribution of
2.4. RESULTS

Figure 2.21: Chromatin modulation affects editing efficiency
(A) Indel profiles at the six selected sites in untreated cells. Indels are ordered by start coordinate relative to the cleavage site (arrowhead), with counts normalised by the effective library size. The mean of two replicates is shown.
(B) Editing efficiency (above) and log$_2$ fold-change in editing efficiency relative to untreated cells (below) for each target site. Biological replicates are shown above and averaged below. NT, untreated cells.
(C) Mean H3K9ac and K3K27ac ChIP-seq (chromatin immunoprecipitation sequencing) signal and DNase-seq (DNase I hypersensitive sites sequencing) signal in a 500 nt window centered on the cleavage site is shown for untreated HepG2 cells (Roadmap Epigenomics Consortium et al., 2015). From Chakrabarti et al. (2019).

We selected six targets with a range of of editing precisions and efficiencies (Figure 2.21A) and individually transfected the corresponding sgRNA into Cas9 expressing cells. We modulated the chromatin using two drugs, both at two doses: i) Trichostatin A (TSA), a histone deacetylase inhibitor, which induced histone hyperacetylation at the target sites and ii) GSK126, an EZH2 inhibitor (EZH2i) which decreased H3K27me3 levels (data not shown).

Treatment with TSA significantly increased editing efficiency in a dose-dependent manner reproducibly across two biological replicates (Figure 2.21B).
Figure 2.22: Effect of chromatin modulation on editing outcome at selected targets

The results for two sites are shown. Above, the counts of individual indels in untreated cells in two biological replicates. Counts are normalised to the effective library size for each replicate; only indels with a normalised count of at least 1 are included. Below, the \( \log_2 \) fold-change in editing efficiency relative to untreated cells. Indel nomenclature is [start coordinate relative to cleavage site]:[size][insertion or deletion]. From Chakrabarti et al. (2019).

For the ACTL6A.5 site, there was nearly a 2-fold increase in efficiency. This effect was most pronounced in target sites that had the lowest starting H3 acetylation levels as determined by ChIP-seq data from the Roadmaps Epigenomics Consortium (Roadmap Epigenomics Consortium et al., 2015) and were also the least accessible as determined using DNase-seq data from the same data set (Figure 2.21C). The relationship with baseline H3 state suggests a direct effect of chromatin modulation on indel formation. EZH2i had the opposite effect: impairing editing efficiency, albeit to a much lesser extent.

Focusing on the specific indel profiles at each site showed that these changes were not restricted to particular indels, but rather they were globally affected by chromatin modulation (Figure 2.22). This was more apparent for the more precise targets such as ACTL6A.5 and BRD2.7 rather than the imprecise targets such as MSH6.2 and SMARCD2.1 (data not shown, Chakrabarti et al. (2019)).
Figure 2.23: Chromatin modulation results in small changes in indel profiles
(A) The normalised relative indel frequency for selected targets in untreated cells (grey bars) and in cells treated with TSA (red outlined bars). The ten commonest indels are shown for each target. NT, untreated cells. Indel nomenclature is [start coordinate relative to cleavage site]:[size][insertion or deletion].
(B) Log\(_2\) fold-change in the relative normalised indel frequency for selected targets relative to untreated cells.
(C) The change in frequency for the three commonest indels for all six targets. The line indicates the mean of both replicates. From Chakrabarti et al. (2019).
2.4.10 Chromatin states influence indel profiles

Although all indels at a site were affected by chromatin modulation, particularly TSA, shorter and longer indels appeared differentially altered (data not shown, Chakrabarti et al. (2019)). Moreover, closer inspection of particular indels at a target site showed some to be preferentially affected, for example –10:12D and –11:11D for ACTL6A.5 (Figure 2.22). To examine these effects of chromatin modulation on indel profiles, we assessed relative changes in the abundance of individual indels, in effect normalising for efficiency. We focused on TSA as the observed changes were greater and more robust, which showed that there were dose-dependent changes in the relative frequencies of indels, with some being favoured more than others (Figure 2.23).  

There were different effects on the commonest indel preference: for MBD3L1.6, MSH6.2 and SMARCD2.1, this was enhanced, but for ACTL6A.5, ASF1B.7 and BRD2.7 it was decreased (Figure 2.23). For the imprecise BRD2.7 site, chromatin modulation altered the commonest indel preference, whereas for the precise ACTL6A.5, the commonest indel remained unchanged despite significant changes to its frequency (Figure 2.23). This suggested that there were differential effects depending on the precision of a target site: indel profiles at precise targets are more robust to changes or differences in chromatin state (for example, across cell types); but indel profiles at imprecise targets are less resilient and the dominant, or most frequent indel can vary.

To complement this experimental modulation of chromatin, we leveraged the fact that the van Overbeek et al. (2016) data set targeted the same 96 sites in three different cell lines (Figure 2.24). We excluded HCT116 from this analysis as this cell line carries a deficiency in mismatch repair which could confound the attribution of changes in indel profile to differences in chromatin. As expected, given our findings of the importance of sequence, indel profiles were broadly similar across the HEK293 and K562 cell lines. However, imprecise sites which had significant differences in accessibility and H3 acetylation also showed differences in their
indel profiles. Regardless, precise targets with such chromatin differences showed preserved, similar indel profiles. This echoed the earlier observations from our own data: precise targets display conserved and reproducible editing outcomes.

2.5 DISCUSSION

2.5.1 Precision of editing outcome

Our ability to predict editing accuracy, efficacy and outcome at specific sites is limited. One of the major obstacles in understanding the rules that guide genome editing, and in generating accurate predictions, is our incomplete understanding.
of how the prokaryotic nuclease interacts with the eukaryotic genome. Eukaryotic genomes are much more complex than their prokaryotic counterparts, for example, with the presence of multiple cellular pathways that recognise and repair DSBs and the packaging of DNA into chromatin. While a number of studies have advanced our understanding of some of these interactions (Brinkman et al., 2018; Isaac et al., 2016; Jensen et al., 2017; Kosicki et al., 2018; Lemos et al., 2018; van Overbeek et al., 2016), the generalisability of their observations is not straightforward due to the limited number of target sites that have been characterised. Here, we have systematically analysed indel profiles at over 1,000 different sites in the human genome, and identified general trends of CRISPR-Cas9 editing and derive simple rules to predict repair outcomes at a given target.

The heterogenous distribution of indel profiles are reproducible and dependent on the local sequence content (which is reflected in the sgRNA and protospacer sequence) (Koike-Yusa et al., 2014; van Overbeek et al., 2016). Here, we extended this observation and found that there is a second layer of heterogeneity on top of these distributions: some targets show one highly preferred indel, while others display a wide range of infrequent, but non-random editing outcomes. We termed this precision. Furthermore, we have shown that editing precision is an intrinsic feature of the target site and depends on 4 key nucleotides that comprise the precision core. Of these 4, the nucleotide at position –4 from the PAM sequence is the most influential. Remarkably, the presence of a “T” at the –4 position gives a site a 51% probability of repairing in a predictable manner and a 91% probability of introducing an insertion.

Our finding that precision of editing is site-specific, and moreover that it can be predicted, has important implications, both for research and clinical applications. Knowing what editing outcome is likely to occur at a given site maximises the chance of generating the desired sequence alteration. Although pharmacological modulation of repair pathways does alter indel profiles, the changes are subtle and for many applications use of such inhibitors may not be suitable (van Overbeek et al., 2016). Selecting a precise site within a gene of interest would be a simpler, safer and more effective way of guiding CRISPR-Cas9 editing to a desired outcome.
2.5. DISCUSSION

Moreover, given the very high reproducibility of indel profiles, selecting a precise target has the potential to edit reliably in a clinical context given the limitations of current alternatives. Both template guided HR and NHEJ have been exploited to is used to knock-in sequences to a target site, but generally these approaches have low efficiencies, particularly in non-dividing cells (Sander and Joung, 2014; Suzuki et al., 2016). Pairs of sgRNA have been used to ‘resect’ the sequence between two target sites, but this can lead to complex, unwanted rearrangements or large on target deletions with pathological consequences (Kosicki et al., 2018).

Although most work to date studying genetic variants and disease has focused on single nucleotide polymorphisms or single nucleotide variants, many diseases are associated with small insertions or deletions, which could be amenable to precise gene editing. Indeed, as a proof-of-concept, template-free Cas9-mediated MMEJ has now been shown to correct pathogenic microduplication alleles endogenously in patient-derived fibroblasts in the HPS1 gene for Hermansky-Pudlak syndrome and in the ATP7A gene for Menkes disease at 88 to 94% efficiency (Shen et al., 2018).

2.5.2 Different routes to precision: complementary studies

Two other studies were published contemporaneously to ours that independently arrived at the concept of precision and the importance of local sequence content (Allen et al., 2018; Shen et al., 2018). Interestingly, both studies took a different approach to ours to address the limitation of needing to characterise indel profiles at scale. They introduced constructs that contained an sgRNA expression cassette as well as the corresponding 20 nt target site with its 3 nt PAM and a larger variable surrounding context (to a total of 55 nt (Shen et al., 2018) or 79 nt (Allen et al., 2018)) into Cas9 expressing human cell lines. Using these constructs they surveyed editing outcomes at 1,872 representative, but synthetic, sites in Shen et al. (2018) and 41,362 in Allen et al. (2018). Our approach was more limited in scale and resolution of the indel profile distribution, but has the advantage of targeting endogenous sites in their native states, with trans-acting factors in situ and the correct nuclear
environment and chromatin context.

Unlike us, both these studies aimed to derive a robust predictive model of editing outcomes that could be used as a tool to guide sgRNA design. Reassuringly, dissecting the InDelphi machine learning model (Shen et al., 2018) and the FORECasT multi-class logistic regression model (Allen et al., 2018) corroborates our findings of the particular importance of the nucleotide at position –4, with smaller contributions from –1, –2, –3 and –5 (Shen et al., 2018). In particular, the remarkable propensity for at “T” at –4 to result in an insertion was recapitulated.

2.5.3 The relationship between precision and indel type

We found a strong correlation between precision of editing and a preference for repairing DSBs with insertions. Targets with an “A” or a “T” at position –4 mainly showed insertions with the commonest insertion occurring at a high frequency and representing on average approximately half of the indels detected at precise site. It is possible that DSB repair with insertions may be kinetically faster compared to other types of indel, and this may partly explain the higher editing efficiency of precise targets and the bias towards single nucleotide indels. However, similar results have now also been obtained by other studies using complementary experimental approaches (Allen et al., 2018; Shen et al., 2018; Taheri-Ghahtarokhi et al., 2018). Notably, the inserted nucleotide can be predicted, as it is nearly always homologous to the nucleotide in the –4 position. Practically, choosing a locus within a gene where an in-frame “TA” dinucleotide is present at positions –5 and –4, could permit the efficient and reproducible introduction of a stop codon and ultimately lead to knock-out of that gene. In contrast, targets with a “G” at position –4 are mostly imprecise and repair induces a variety of unpredictable deletions.
2.5.4 Nucleotide –4 plays a key role in defining site-specific indel profiles

A subsequent study using a similar approach to ours, surveying 1656 endogenous sites, has also validated our findings with regards to the nucleotide at position –4 and insertions and deletions: “T” and “A” tend to promote insertions, while “C” and “G” tend to promote deletions (Leenay et al., 2019). It is particularly interesting to speculate on the mechanistic link between the nucleotide at position –4 and its effect on defining editing precision and preferences for indel types. Cas9 was predominantly thought to cleave DNA to produce blunt ends, however, there is a small but growing body of evidence, that in fact it can produce staggered ends (Shou et al., 2018; Zuo and Liu, 2016). The generation of 5′ overhangs are mostly observed at position –4 on the non-complementary strand. These results can explain both the prevalence of single nucleotide insertions and the insertion homology we observed: the overhanging nucleotide is used as a template for polymerase θ before the ends are rejoined. So, paradoxically imprecision of Cas9 cleavage is the likely cause of precision in the insertion outcome. The high frequency of single nucleotide deletions could also be accounted for by the staggered cleavage, if the Ku proteins are less able to bind the ragged ends, and this may permit 1 nt end processing back to two blunt ends, which are then ligated together.

How the base composition of the –4 position may influence editing precision and outcomes is not clear. In the case of precise homologous insertions favoured by a “T” or an “A” it may be that these have varying abilities to recruit proteins involved in DNA repair, notably DNA polymerase. Alternatively, more likely possibility is that Cas9 is more likely to make an asymmetric cut with these nucleotides in position –4. The other nucleotides in the precision core could act through similar mechanisms.

We propose a model where flexible cleavage by Cas9 influences DNA repair fidelity: when blunt ends are generated at position –3, cells repair the DSB in an error free manner reconstituting the original sequence, whereas indels mainly occur when assymetric cleavage generates staggered ends. This model is based on our
observation that the vast majority of insertions show homology, and other findings that NHEJ-mediated repair of CRISPR-Cas9 induced DSBs is mostly error-free (Geisinger et al., 2016) and that deletions generated by sgRNA pairs can be repaired with a high level of precision (Shou et al., 2018). This model may reconcile apparently conflicting results related to the fidelity of NHEJ in CRISPR-Cas9 dependent and independent scenarios (Brinkman et al., 2018; Dudley et al., 2005; Geisinger et al., 2016; Shou et al., 2018; van Heemst et al., 2004). Both outcomes can be exploited for gene editing: blunt ends could allow precise insertions of exogenous sequences (Suzuki et al., 2016) while staggered ends enable the induction of indels to knock out genes.

2.5.5 The influence of the chromatin environment on site-specific editing outcomes

Local DNA sequence context is the major determinant of precision and site-specific indel profiles. However, we have shown that the chromatin context of the target site does also play an ancillary role. The packaging of DNA into chromatin can affect both the editing efficiency and the relative frequency of indels at a given locus. We found that increasing histone hyperacetylation, which enables a more open chromatin state, enhanced indel formation.

Although the effect of the histone deacetylase inhibitor, TSA, was observed at all sites, it was particularly pronounced at poorly accessible sites, as assessed by DNase-seq, with low endogenous levels of histone acetylation. The results across all six sites were suggestive of a dose-response relationship. Transient treatment with TSA could be used as a strategy to boost editing at sites that are located in more repressive, closed chromatin states. These results are consistent with in vivo studies showing positive correlations between sgRNA activity and open chromatin at genome-wide levels (Uusi-Mäkelä et al., 2018) and in vitro studies showing that nucleosome positioning at a target site can impede the binding of Cas9 to DNA and inhibit its activity (Horlbeck et al., 2016b).

We have also shown that modulation of chromatin has differential effects on
the individual indels at a target site. While these are often subtle, albeit measurable, they can result in a change in the identity of the commonest indel at more imprecise targets. It is notable that the magnitude of these changes are comparable with those observed after inhibition of the NHEJ repair pathway (van Overbeek et al., 2016). Chromatin has an established role in DNA repair (Kalouzi and Soutoglou, 2016) and multiple DNA repair pathways are involved in CRISPR-Cas9 editing (Brinkman et al., 2014; Maruyama et al., 2015; Shou et al., 2018; van Overbeek et al., 2016), and so variable recruitment of factors involved in these pathways as a result of the either the chromatin environment around the target locus, or modulation of this environment may result in the differences we have observed. However, it is important to reiterate that these alterations only apply to imprecise targets; precise targets show consistent dominant indels regardless of chromatin states with preserved editing outcomes across cell types.
# 3 Evaluation of Methods to Study Protein-RNA Interactions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Summary of the chapter</td>
<td>100</td>
</tr>
<tr>
<td>3.2</td>
<td>CLIP technologies to study protein-RNA interactions</td>
<td>100</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Experimental factors affecting CLIP data quality</td>
<td>101</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Considerations for CLIP data analysis</td>
<td>107</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Computational quality control of CLIP data</td>
<td>107</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Peak calling approaches for CLIP data</td>
<td>111</td>
</tr>
<tr>
<td>3.2.5</td>
<td>An evaluation of CLIP methods and analysis approaches</td>
<td>113</td>
</tr>
<tr>
<td>3.3</td>
<td>Methods</td>
<td>114</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Processing of CLIP data</td>
<td>114</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Processing of PTBP1 knockdown splicing microarray data</td>
<td>118</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Processing of PTBP1 knockout RNA-seq data</td>
<td>118</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Generating RNA maps</td>
<td>119</td>
</tr>
<tr>
<td>3.4</td>
<td>Results</td>
<td>120</td>
</tr>
<tr>
<td>3.4.1</td>
<td>An overview of the data sets</td>
<td>120</td>
</tr>
<tr>
<td>3.4.2</td>
<td>RNA maps to compare PTBP1 CLIP data</td>
<td>122</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Assessing the raw CLIP data quality</td>
<td>123</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Assessing the effect of peak calling on data quality</td>
<td>125</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Assessing the effect of different peak callers on data quality</td>
<td>127</td>
</tr>
<tr>
<td>3.4.6</td>
<td>Assessing optimisation of peak calling parameters</td>
<td>127</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusions</td>
<td>132</td>
</tr>
</tbody>
</table>
CHAPTER 3. EVALUATION OF METHODS TO STUDY PROTEIN-RNA INTERACTIONS

3.1 SUMMARY OF THE CHAPTER

This chapter tackles the analysis of genomic data from experimental approaches that are used to study RNA-protein interactions. I start by introducing UV crosslinking and immunoprecipitation (CLIP) technologies that have been developed to identify direct endogenous protein-RNA interactions. I focus on the computational analysis of these data, relating this back to the experimental method that has been used to generate them. I discuss data quality: the factors that influence it, and how to assess it. In particular, I demonstrate the utility of visualising RNA maps, which examine the positional distribution of binding around regulated landmarks in transcripts. I use these RNA maps to show how variations in different CLIP methods and in different peak calling methods and parameters affect data quality and insights into regulatory mechanisms.

Parts of section 3.2 and 3.4 have been published and are adapted from Chakrabarti et al. (2018) with new additional analyses and insights. The code to plot RNA maps was developed from the original version written by Dr Nejc Haberman.

3.2 CLIP TECHNOLOGIES TO STUDY PROTEIN-RNA INTERACTIONS

RBPs are key orchestrators of post-transcriptional regulation of RNAs as discussed in the Introduction. CLIP methods are the mainstay of direct examination of RBP-RNA interactions in vivo. There are over 30 published protocols with iterative improvements or changes to the original CLIP method (Lee and Ule, 2018; Ule et al., 2003). However, the 11 core steps remain largely unaltered (Lee and Ule, 2018) and it is crucial to understand the nuances to ensure appropriate computational analyses of the resultant sequencing data (Chakrabarti et al., 2018). The experimental developments and methodology of CLIP has been extensively reviewed in Lee and Ule (2018). Here I distil the salient general steps that it is necessary to appreciate in order to guide the appropriate computational analysis of the sequencing data.
(Figure 3.1, steps 1–11 and Table 3.1).

Irradiation with UVC at 254 nm crosslinks RNAs and proteins at ‘zero distance’ due to the photoreactivity of the RNA bases (Shetlar et al., 1984). This allows fragmentation of the RNAs using a limited concentration of RNase after cell lysis. The RBP-of-interest can then be purified under stringent conditions, usually through the use of a specific antibody against the RBP. The protein-RNA complexes are then separated by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) which denatures and isolates proteins by their molecular weight. A very important quality control step is the visualisation of the protein in complex with the crosslinked RNA fragments. The section of the membrane containing the protein-RNA complex is then excised and treated with proteinase to remove the bulk of the RBP, leaving behind only a short polypeptide that remains attached at the crosslink site. This also releases the RNA fragments that can then be reverse transcribed into complementary DNAs (cDNAs). This library can then be subjected to high-throughput sequencing for a transcriptome-wide view of the binding patterns of the RBP.

3.2.1 Experimental factors affecting CLIP data quality

3.2.1.1 Resolution and sensitivity

CLIP methods fall into three main categories based on the effect on reverse transcription of the polypeptide that remains covalently bound at the crosslink site of fragmented RNAs (Table 3.1). These diagnostic events determine how the sequencing data should be approached for data analysis and largely also the sensitivities of the different methods (Chakrabarti et al., 2018; Drewe-Boss et al., 2018).

1. cDNAs that read through the peptide without any mutations
2. cDNAs that read through the peptide but introduce a mutation at the crosslink site
3. cDNAs that truncate at the crosslink site

In the original CLIP and HITS-CLIP (high throughput sequencing CLIP) meth-
CHAPTER 3. EVALUATION OF METHODS TO STUDY PROTEIN-RNA INTERACTIONS

Figure 3.1: An overview of the iCLIP method and its analysis. From Chakrabarti et al. (2018), with permission by Annual Reviews.
### Methods, Resolution, Specificity, Sensitivity

<table>
<thead>
<tr>
<th>Methods</th>
<th>Resolution</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>HITS-CLIP, CLIP-seq</td>
<td>Oligonucleotide corresponding to size readthrough cDNAs</td>
<td>++ to +++</td>
<td>++</td>
</tr>
<tr>
<td>iCLIP, irCLIP</td>
<td>Nucleotide corresponding to the start of truncated cDNAs</td>
<td>++ to +++</td>
<td>++ to +++</td>
</tr>
<tr>
<td>eCLIP</td>
<td>Nucleotide corresponding to the start of truncated cDNAs</td>
<td>+ to +++</td>
<td>++ to +++</td>
</tr>
<tr>
<td>PAR-CLIP</td>
<td>Nucleotide corresponding to the crosslink induced mutations</td>
<td>++ to +++</td>
<td>+ to +++</td>
</tr>
<tr>
<td>CIMS of HITS-CLIP</td>
<td>Nucleotide corresponding to the crosslink induced mutations</td>
<td>++ to +++</td>
<td>+</td>
</tr>
<tr>
<td>RIP-seq</td>
<td>Transcript-level since it does not fragment the bound RNAs</td>
<td>+</td>
<td>++</td>
</tr>
</tbody>
</table>

*Table 3.1: The central features of CLIP from the perspective of data analysis.*

CLIP methods are grouped according to how the reads are used to identify binding sites (indicated in the resolution column). +, moderate; ++, high; +++; best. CLIP, UV crosslinking and immunoprecipitation; HITS-CLIP, high-throughput sequencing CLIP; iCLIP, individual-nucleotide resolution CLIP; irCLIP, infrared CLIP; eCLIP, enhanced CLIP; PAR-CLIP, photoactivatable ribonucleoside-enhanced CLIP; CIMS, crosslink-induced mutation sites; RIP, RNA immunoprecipitation. Adapted from Chakrabarti et al. (2018), with permission by Annual Reviews.
ods both adapters needed for cDNA amplification were ligated to the RNA fragments, resulting in the full length fragment being reverse transcribed (Licatalosi et al., 2008; Ule et al., 2003). Consequently, only reads falling into categories 1 and 2 can be used. This means that binding sites can only be assigned at oligonucleotide resolution on the basis of the whole read. The use of UVC light usually only leads to a small proportion of mutations in the cDNAs (Sugimoto et al., 2012). The particular reverse transcriptase that is used also has implications for the rate and type of mutation that is induced (Van Nostrand et al., 2017b). Nevertheless, crosslink-induced mutation sites (CIMS) have been used to increase the resolution of HITS-CLIP to nucleotide-level by identifying the positions of these mutations within the read (Zhang and Darnell, 2011).

Photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) increased the proportion of mutations that are introduced in the cDNA at the crosslink sites by pre-incubating cells with 4-thiouridine (4sU). These bases can be crosslinked to proteins with UVA irradiation at 365 nm (Hafner et al., 2010). As for CLIP and HITS-CLIP, both readthrough and mutation reads (i.e. categories 1 and 2) are amplified, but with PAR-CLIP, approximately 50% of the cDNAs contain T-to-C transitions at the crosslink site. This is the basis for assigning binding sites to nucleotide resolution using PAR-CLIP sequencing data (Corcoran et al., 2011; Hafner et al., 2010). However, this does mean that about 50% of cDNAs do not contain transitions, and longer cDNA may contain multiple transitions; both of these types of cDNAs are discarded to enable nucleotide resolution analysis.

Individual-nucleotide resolution CLIP (iCLIP) was developed in particular to isolate the third category: truncated cDNAs (König et al., 2010). The key innovation enabling this was ligating the first adapter to the RNA, but the second adapter to the cDNA, after the reverse transcription has truncated prematurely at the crosslink site (König et al., 2010; Lee and Ule, 2018). Studies have estimated that approximately 90% of cDNAs truncate at the crosslink site in iCLIP (Haberman et al., 2017; Sugimoto et al., 2012). This means that the adapter is ligated exactly at the position of truncation in the vast majority of cases. Hence, the nucleotide position in the genome immediately 5′ to the mapped cDNA nearly always corresponds to the crosslink
3.2. CLIP TECHNOLOGIES TO STUDY PROTEIN-RNA INTERACTIONS

Figure 3.2: A comparison of iCLIP, eCLIP and irCLIP experimental methods.

The core steps are categorised for iCLIP as in Lee and Ule (2018). From the top:
1. Covalent protein-RNA crosslinking; 2. Cell lysis; 3. RNA fragmentation; 4. Purification of protein-RNA complexes; 5. Ligation of SeqRv adapter to fragmented RNA; 6. Quality control; 7. RNA extraction; 8. Reverse transcription; 9. cDNA purification; 10. SeqFw adapter ligation; 11. cDNA amplification and sequencing. Next to iCLIP, key differences in the eCLIP and irCLIP methods at particular steps are indicated. (Adapted from figures by Dr Flora Lee.)
position. Truncation-based methods have an inherently greater sensitivity as they can maximise the capture of crosslinking events. Analysis of truncated cDNAs forms the mainstay of the majority of CLIP data analysis as it is applicable also to the more recent iCLIP derivatives irCLIP (infrared CLIP) (Zarnegar et al., 2016) and eCLIP (enhanced CLIP) (Van Nostrand et al., 2016). The latter has been used by the ENCODE consortium to study hundreds of different RBPs (Van Nostrand et al., 2017a). The core experimental steps of iCLIP, and the differences in eCLIP and irCLIP are shown in Figure 3.2.

3.2.1.2 Specificity

Ensuring the specificity of CLIP (that only RNA fragments bound to the protein of interest are isolated and sequenced) largely depends on the experimental strategies adopted. These include immunoprecipitation (IP) under stringent conditions, as in iCLIP (Huppertz et al., 2014; König et al., 2010) or under denaturing conditions with epitope tags as in cross-linking and analysis of cDNAs (CRAC) (Granneman et al., 2009). Validation of the proteins within the isolated complex is sometimes also confirmed with mass spectrometry. This can be of particular importance when considering more stable RNPs that contain multiple RBPs. However, when studying the complex as a whole, varying degrees of immunoprecipitation stringency can isolate different parts of the complex and be leveraged to study their roles, for example with spliceosome iCLIP where harsh conditions isolate only the SmB protein; moderate conditions, the snRNPs; and mild conditions the multi-protein spliceosomal complex (Briese et al., 2018).

In the absence of a denaturation step, the purification of the immunoprecipitated protein-RNA complexes is crucial to ensure specificity and remove RNA that is either unbound or potentially bound to other RBPs. In iCLIP, irCLIP and eCLIP, SDS-PAGE and membrane transfer is used. Moreover, it is as important for quality control to visualise these complexes to control for the specificity of the purified complexes. In the original iCLIP this was performed with radioactive 5’ end labelling of the RNA fragment (Huppertz et al., 2014; König et al., 2010), while in irCLIP an infrared signal is introduced via a dye-coupled adapter (Zarnegar et al., 2016). In
3.2. CLIP TECHNOLOGIES TO STUDY PROTEIN-RNA INTERACTIONS

methods where this quality control step is omitted, such as eCLIP, the data specificity cannot be assured and so additional assessment needs to be undertaken during the computational analysis as explored later (Chakrabarti et al., 2018). In eCLIP a parallel sample is processed with a mock immunoprecipitation step without an antibody: hence all the crosslinked RNAs are processed as for the eCLIP with the immunoprecipitation, and this sample is used to judge background signal as it contains ‘all’ of the RNAs bound by RBPs of a similar size to the RBP under study.

3.2.2 Considerations for CLIP data analysis

Next, I summarise the main computational steps of the CLIP method (Figure 3.1, steps 12 - 16) as they apply to the truncation-based CLIP methods that are evaluated later. I then consider in more detail two key aspects important for this chapter: data quality control and peak calling. Other aspects, such as motif discovery and the modelling of binding sites to address the false negative problem have also been explored further in Chakrabarti et al. (2018).

First the sequenced reads are aligned, usually to the genome, taking care to ensure the start of the read is preserved during alignment and not soft-clipped. This is necessary to ensure the correct assignment of the crosslink site as the position immediately 5′ to the read start. Then the crosslink sites are used to call peaks to identify those sites that are highly bound by an RBP and therefore most likely to be functionally significant. Here, it is important to account for variable RNA abundance as this will influence the read counts on a given transcript. Finally, the binding site peaks can be used to integrate the CLIP data with orthogonal data, both to assess data quality (as done in Section 3.4) and to derive biological insight.

3.2.3 Computational quality control of CLIP data

The assessment of data quality is essential to ensure validity of the results from a CLIP experiment. Both sensitivity and specificity of the data, which the experimental
techniques have aimed to optimise, need to be assessed together.

3.2.3.1 cDNA complexity

The simplest measure of sensitivity is the total number of unique cDNAs in the library. This cDNA complexity enables an estimation of the dynamic range of the experiment and of the RBP-RNA interactions that can be detected. A number of factors, both biological and technical, can affect cDNA complexity including RBP abundance and crosslinking efficiency, immunoprecipitation efficiency, adapter ligation and cDNA library preparation. In particular, PCR duplication not only has the potential to confound biological results, but also hampers accurate monitoring of complexity as the amplification of cDNA fragments is not uniform but affected both by sequence content and length. Optimal handling of the removal of PCR duplicates (Figure 3.1, step 13) is therefore important when assessing cDNA complexity and sensitivity. Some approaches simply collapse identical sequences before alignment (Flynn et al., 2015; Shah et al., 2017; Zarnegar et al., 2016), while others refine this by including positional information based on the aligned start positions of the reads to account for read variation arising due to sequencing errors (Zhang and Darnell, 2011). However, the current gold-standard is the use of unique molecular identifiers (UMIs): a random sequence of nucleotides (commonly 5-10 nt) is introduced experimentally to label each cDNA as it is reverse transcribed. After amplification, the UMI can be used to trace the sequenced reads back to the unique cDNAs in combination with aligned read start information (Curk et al., 2017; König et al., 2010; Smith et al., 2017). While all of these improve the accuracy of the total unique cDNA count estimate, this is still a relatively crude sensitivity estimate. A more refined measure of sensitivity using the crosslink positions obtained from unique cDNAs is through the use of RNA maps (Section 3.2.3.3), where a functionally relevant assessment can be made.

3.2.3.2 cDNA specificity

Assessing cDNA specificity during the computational analysis can be challenging for many RBPs, as often the ground truth of where the RBP binds is unknown.
Consequently, often only circumspect estimations can be made. A basic measure is the percentage of crosslink sites that occur within peaks as it gauges the capacity of the library to identify binding sites. A complementary approach is the comparison of the enrichment of RBP-specific \( k \)-mers or motifs within peaks compared to an appropriately defined background region. These can either be defined from the CLIP data, or from alternative methods such as the in vitro RNAcompete assay (Ray et al., 2013). However, this would only work for RBPs that bind with sequence specificity. Moreover, all of these assessments also require appropriate optimisation of peak calling, discussed later. The best measure of specificity is by using RNA maps (Section 3.2.3.3) to visualise CLIP signal with orthogonal functional data about the RBP, where this is available.

3.2.3.3 RNA maps facilitate simultaneous assessments of sensitivity and specificity

RNA maps are a powerful tool to integrate orthogonal data sets to study mechanisms of RNA regulation (Chakrabarti et al., 2018; Witten and Ule, 2011). They have most commonly been used to assess the regulation of splicing by RBPs; in their first use, binding motifs were integrated with microarray splicing data to predict the genome-wide action of NOVA (Ule et al., 2006). They are a conceptually simple analysis that visualises the positional distribution of assigned binding sites around regulated transcript landmarks. Binding sites are usually defined using CLIP data (such as crosslinks or peaks), although CLIP- or in vitro-derived RBP motifs can also be used. The regulated landmarks are usually selected using complementary data and depend on the regulatory mechanism under investigation: for splicing, for example, RNA-seq analysis of RBP knockout cells or tissues can identify the regulated exons. Typically this distribution is summarised as a metaprofile around a hypothetical combined landmark (Chakrabarti et al., 2018; Park et al., 2016a; Witten and Ule, 2011; Yee et al., 2019), although heatmaps are also used for more detailed investigation (Chakrabarti et al., 2018; Haberman et al., 2017; Rot et al., 2017).

A schematic of an RNA map is shown in Figure 3.3 for assessment of an RBP that regulates splicing, with the CLIP crosslink distributions shown around the 3′ and 5′ splices sites of enhanced and silenced meta-cassette exons. It is important
EVALUATION OF METHODS TO STUDY PROTEIN-RNA INTERACTIONS

Figure 3.3: Anatomy of an RNA map
A schematic of an RNA map showing CLIP binding around regulated spliced exons. The 3′ splice site is shown on the left and the 5′ splice site on the right. The region depicting the exon is shaded in grey. The blue line shows a peak of binding upstream of the 3′ site for silenced exons, while the red line shows distributed binding downstream of the 5′ splice site for enhanced exons. The grey line shows binding around control exons. Crosslink coverage means the proportion of regulated exons with crosslinks covering at a given position. For the PTBP1 RNA maps presented in this chapter, the sensitivity score is calculated as the height of the silenced peak: $h_s$, and the specificity score as the ratio of the silenced peak to the maximum enhanced coverage in the –100 to 0 window: $\frac{h_s}{h_E}$. The specifics of these calculations would be dependent on the RBP under study and its particular binding patterns.

to use a control set of unregulated exons to determine enrichment (either through visual assessment or calculation).

In addition to their use in deriving biological insight, RNA maps are also useful for assessing data quality and validating CLIP experiments. This facet is exploited in Section 3.4 to compare different experimental and computational methods. The proportion of regulated RNAs (cassette exons in the schematic) with CLIP peaks at the expected position provides an assessment of sensitivity. Meanwhile, the enrichment of CLIP peaks around these RNAs affords a measure of specificity.
3.2.4 Peak calling approaches for CLIP data

For CLIP data, the primary aim of peak calling is to identify sites that are highly occupied by an RBP and are therefore most likely to be of functional relevance. This is in contrast to ChIP-seq data where the primary aim is to remove background signal and noise. Although, peak calling in CLIP can also achieve this, the high specificity of a good quality CLIP experiment, with its multiple high-stringency purification steps, ensures that the cDNAs that are sequenced are truly reflective of RBP-bound RNA fragments. RBP binding cannot be classified into simple binary categories of specific and non-specific. Rather there is a gradation of binding affinities, with different modes of binding. Some RBPs, for example, transiently associate with multiple low-affinity sites of nascent transcripts through association with RNA polymerase, before finding a high-affinity site (XRN2, for example). Others bind to a high-affinity site and then spread over larger regions (PTBP1, for example). Hence, although an RBP will likely crosslink repetitively to a clustered set of crosslinks around a high-affinity site, there are no absolute thresholds to use to differentiate these sites from the lower affinity ones, as this depends on a wide range of other factors (Chakrabarti et al., 2018). This is why it is paramount to ensure the data are of high quality and specificity prior to peak calling to derive this more nuanced characterisation of RBP binding.

The basic approach of peak calling is to search for pileup of aligned reads or diagnostic events that correspond to crosslinks. Many peak calling tools have been developed to refine this strategy (Chakrabarti et al., 2018; De and Gorospe, 2017; Wheeler et al., 2018). There are five main considerations faced when undertaking peak calling:

1. What to use to call a peak?
2. How to define a peak?
3. How to account for variable RNA abundance?
4. How to account for crosslinking biases?
5. How reproducible are the data?
Here I elaborate on how 4 different peak calling tools: iCount (Curk et al., 2017), Piranha (Uren et al., 2012), CLIPper (Lovci et al., 2013) and Paraclu (Frith et al., 2008) address the second question. The first three have been developed for CLIP data, while the last was developed for transcription start site identification from CAGE (cap analysis of gene expression) data. Other tools and the other considerations have been previously discussed in detail (Chakrabarti et al., 2018).

All the user-defined parameters of the various tools require some understanding of the RBP under study and its binding pattern or preference to optimise peak calling. The iCount, Piranha and CLIPper algorithms are all based on calculating the probability that a given crosslink site does not belong to a background CLIP read distribution. Adjacent significant crosslink sites are then grouped to form peaks. The differences in these tools largely stem from which probability distribution function they use to model the counts and how they generate the background distribution, and how they define the boundaries of a peak.

Piranha (Uren et al., 2012) calculates crosslink counts across the whole genome in user-defined bin sizes and assumes most bins contain background noise. A zero-truncated negative binomial distribution is fitted to the data. This accounts for overdispersion in the count data and has been shown to fit CLIP data better than simple negative binomial or Poisson distributions (Uren et al., 2012). Bins where there is a higher read count than would be expected are deemed significant and can be selected using a p-value threshold. Bins within a user-defined distance are clustered to form peaks.

iCount (Curk et al., 2017) uses permutation analysis rather than a specific probability distribution. The crosslinks are randomly distributed a user-defined number of times within a relevant region of interest (e.g. introns) on a gene-by-gene basis, to generate a background distribution for that gene. The comparison of the observed distribution (where crosslink counts for a given position are weighted by all the counts within a user-defined window centred on the position) with the random one generates a false discovery rate. Significant crosslink sites can be selected using a false discovery rate threshold and are also clustered into peaks using a user-defined distance.
3.2. CLIP TECHNOLOGIES TO STUDY PROTEIN-RNA INTERACTIONS

CLIPper (Lovci et al., 2013) combines aspects from both these tools. Similar to iCount, a false discovery rate is calculated during a first pass. However, the crosslinks are randomised over the whole gene by default, rather than on a regional basis. In a second pass, similar to Piranha, albeit using a Poisson distribution, crosslink sites that have fewer counts than would be expected across the transcriptome are removed. Finally cubic-spline fitting is used to fit a curve to the significant crosslink sites to form a peak with the boundaries defined by excluding points that exceed the false discovery rate.

Paraclu (Frith et al., 2008) was developed for a different type of data, and so uses a different strategy. In essence it is a clustering tool that imposes minimal assumptions on the data. Applied to CLIP, the algorithm reports segments of the genome that maximise a density score based on the number of crosslinks in a segment and the segment size. It favours segments with a large number of events, but disfavours large segments, so the reported segments are the ones with the best compromise between these two factors and are effectively prominent clusters of observed crosslinks. As this reports clusters of multiple sizes, i.e. clusters within clusters, it is necessary to set a user-defined threshold for the maximum cluster size to define the appropriate boundaries.

3.2.5 An evaluation of CLIP methods and analysis approaches

In order to get a better understanding on how different CLIP methods affect data quality and the biological conclusions that may be drawn, I focus on the RBP PTBP1. This is a well-characterised RBP (Attig et al., 2018; Cereda et al., 2014; Haberman et al., 2017; Llorian et al., 2010; Xue et al., 2009) that predominantly binds introns to regulate splicing (Cereda et al., 2014; Haberman et al., 2017; Xue et al., 2009). In vitro RNAcompete binding assays have identified a “HYUUUYU” binding motif (Ray et al., 2013) and in vivo motif analysis of iCLIP data identified a CU-rich multivalent motif (Cereda et al., 2014). The high uridine content of its motif means that PTBP1 is an RBP that crosslinks well, on account of a moderate bias of UVC crosslinking to
U-rich motifs (Sugimoto et al., 2012). PTBP1 contains four RRM domains that all bind and sequester pre-mRNA regions to promote exon skipping, with RRM3 and RRM4 binding to separated sites and looping out the intervening RNA (Mackereth and Sattler, 2012). This mode of binding, with multiple RRMs acting in concert, is similar to other splicing factors. Furthermore, PTBP1 has been studied using iCLIP, eCLIP and irCLIP with representative data produced by the laboratories that developed these CLIP-variant methods.

All three of these methods have nucleotide resolution and focus on identifying truncated cDNAs. The other widely used method, PAR-CLIP (Hafner et al., 2010; Mukherjee et al., 2019) uses mutations, enriched in frequency by 4sU treatment, to identify crosslink sites. Hence a different computational approach is necessary to identify the diagnostic crosslink events reliably over background mutations and sequencing errors. So, in order to maintain as equitable a computational approach as possible I have not considered PAR-CLIP in my comparative analysis.

Given the explosion in different variants of CLIP technologies (Lee and Ule, 2018), and in different analysis approaches (Chakrabarti et al., 2018; Uhl et al., 2017) the development of methods to evaluate the data quality and the results they produce is crucial. Here, I assess the strengths and weaknesses commonly used experimental and computational methods to guide future experiments.

3.3 METHODS

3.3.1 Processing of CLIP data

3.3.1.1 Adapter trimming and UMI extraction

On account of the different experimental protocols, slightly different methods were needed to remove adapter sequences and extract the UMIs from the sequenced reads in the FASTQ files.

- For iCLIP, 3’ adapters and the experimental barcodes were removed and the UMI moved to the FASTQ header using iCount with the adapter se-
3.3. METHODS

sequence AGATCGGAAG (Curk et al., 2017). The 5′ barcode has the structure 5′–NNNNNNNNNNNNNNN where N is the UMI and X is the experimental barcode.

- For irCLIP, the 3′ adapter sequence AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTTG was trimmed using Cutadapt (Martin, 2011). The 5′ barcode has the structure 5′–NNNNNNNNN where N is the UMI and X is the experimental barcode. The extracted UMI was placed in the FASTQ header and these 17 nt nucleotides trimmed from the read.

- For eCLIP, the UMI had already been removed and placed in the FASTQ header by the ENCODE consortium. For adapter removal, two rounds of trimming were performed using Cutadapt (Martin, 2011) according to the ENCODE eCLIP standard operating procedure. The first round trims both 5′ and 3′ adapters and the second round 3′ adapter double ligation events.

3.3.1.2 Read alignment

Reads were aligned in a two-stage alignment strategy to maximise the proportion of reads mapping uniquely to the genome. First, reads were aligned using Bowtie2 (Langmead and Salzberg, 2012) to a custom index generated using rRNA subunit sequences from NCBI (5S, NR_023363.1; 5.8S, NR_003285.2; 28S, NR_003287.2; and 18S, NR_003286.2) and tRNA sequences obtained from GtRNAdb (Chan and Lowe, 2009) with default parameters. Unmapped reads were taken forward to the second step. Reads were aligned to the GRCh37 genome assembly with the Ensembl 75 annotation using STAR (Dobin et al., 2013). iCLIP and irCLIP reads were aligned as single-end reads, but eCLIP was aligned as paired-end, on account of the different methods of library generation. iCLIP was mapped with the additional parameters:

--outFilterMultimapNmax 1 --outFilterMultimapScoreRange 1
--outSAMattributes All --outFilterType BySJout
--outFilterScoreMin 10 --alignEndsType EndToEnd
--outFilterMismatchNmax 2

This ensured the 5′ end of the read, corresponding to the cDNA start, was not
soft-clipped during alignment and only uniquely mapped reads were retained. With these parameters, for irCLIP and eCLIP, there was a high proportion of reads unmapped on account of too many mismatches using these stringent mismatch settings which are the default for iCLIP analysis (König et al., 2010; Zarnack et al., 2013) so \texttt{-outFilterMismatchNmax 2} was omitted giving preference to the default STAR parameters. For all data only uniquely mapped reads were kept.

3.3.1.3 **PCR duplicate removal and crosslink file generation**

PCR duplicate removal was performed using a custom R script for iCLIP and irCLIP and using the \texttt{barcode_collapse_pe.py} script from ENCODE. Both of these methods collapsed mapped reads with the same chromosome, start position and UMI to one representative read. BAM files were converted to crosslink BED files with the coordinates shifted to the position immediately 5′ to the start position in a strand-aware manner either within the custom R script for iCLIP and irCLIP, or with custom Python scripts written by Dr Nejc Haberman for eCLIP. For the tiny minority of reads where this results in a position beyond the chromosome, the start position was taken instead.

3.3.1.4 **Regional distribution**

The genomic regions in which crosslinks were located were identified by intersecting the crosslink BED files with a custom annotation BED file. This annotated each genomic coordinate with a unique region and was generated by collapsing the full GRCh37 Ensembl 75 GTF annotation file into a non-redundant annotation based on the following hierarchy: ncRNA > CDS > 3′ UTR > 5′ UTR > intron > other. Any unannotated crosslinks were deemed to be intergenic.

3.3.1.5 **Peak calling**

A number of different peak calling tools and parameters were evaluated.
3.3. METHODS

**iCount** For iCount (Curk et al., 2017) a segment file was first generated using the Ensembl 75 GTF file and the command: iCount segment GRCh37.75.gtf.gz GRCh37.75.seg.gtf.gz GRCh37.fa.fai. This segment file was then used for all iCount peak calling. Summed crosslink BED files were created by collapsing the crosslinks to unique crosslink positions, with the BED score assigned as the total crosslinks at that position. This was either done within the custom R script for PCR duplicate removal and crosslink file generation, or using BEDTools and AWK, with the command: sort -k1,1 -k2,2n -k3,3n -k6,6 <crosslink BED file> | bedtools groupby -i stdin -g 1,2,3,6 -c 5 -o sum | awk '{OFS="\\t"}{print $1,$2,$3,".",$5,$4}'> <summed crosslink BED file>.

There are two steps to iCount peak calling: first, identifying significant crosslink positions and second, clustering significant crosslink positions to form peaks. The first step was carried out using the command: iCount peaks GRCh37.75.seg.gtf.gz <summed crosslink BED file> <significant crosslink BED file> -fdr 0.05 -half_window w, where w is the user-specified half-window size (i.e. how many positions either side of a crosslink position to sum for the permutation analysis for a given position). The second step was carried out using BEDTools using the command bedtools merge -i <significant crosslink BED file> -s -d d -c 4,5,6 -o distinct,sum,distinct > <peak BED file>, with d indicating the clustering distance and the BED score recalculated as the number of crosslinks within a given peak.

**Piranha** For Piranha (Uren et al., 2012), significant crosslinks were identified using the command: Piranha -z w -u 0 -o <significant crosslink BED file> <crosslink BED file>, with w indicating the bin size into which to bin the reads. Merging the significant crosslink file was disabled with -u 0 to permit finer control for clustering the significant crosslink positions to form peaks using the BEDTools command as for iCount.
CHAPTER 3. EVALUATION OF METHODS TO STUDY PROTEIN-RNA INTERACTIONS

CLIPPER For CLIPper (Van Nostrand et al., 2016), the crosslink count BED file was converted to a BAM file using the BEDTools command: `bedtools bedtobam -i <crosslink BED file> -g <chromosome lengths file> > <crosslink BAM file>`. This was used as input to CLIPper, which was run using the command: `clipper -b <crosslink BAM file> -s hg19 -o <peak BED files>`

PARACLU For Paraclu (Frith et al., 2008), there are also two steps to peak calling. First, the summed crosslink BED file was converted to a 4-column file (chromosome, strand, position, count). This was used as the input to Paraclu: `paraclu m <input file> > <intermediate file>` where $m$ is the minimum number of reads to form a cluster. Second, this intermediate output file was then filtered using the command `paraclu-cut.sh -l w -d 2 <intermediate file> > <output Paraclu file>` to omit clusters that are singletons, above a threshold width $w$, or with a fold-increase in density of $<2$ (the default). Finally, the Paraclu output file was converted to a peak BED file. All of these conversions and the running of Paraclu was performed within a custom wrapper R script.

3.3.2 Processing of PTBP1 knockdown splicing microarray data

Processed splicing microarray data upon PTBP1/2 knockdown in HeLa cells were kindly provided by Dr Nejc Haberman using data from accession E-GEOD-23513 (Llorian et al., 2010). Enhanced, silenced and control exons were identified as described in Cereda et al. (2014). In brief, regulated exons were identified as those with a dIrank of greater than 1 using ASPIRE3.

3.3.3 Processing of PTBP1 knockout RNA-seq data

RNA-seq data from K562 cells upon CRISPR knockout of PTBP1 were obtained from ENCODE as BAM files aligned to the GRCh37 genome using the Gen-
code V19 annotation. PTBP1 knockout accession numbers were ENCFF780QXW and ENCFF914MST and control accession numbers were ENCFF925GZR and ENCFF842ZDM. rMATS (Shen et al., 2014) was used to identify alternative splicing events using the command:

```
RNASeq-MATS.py -len 100 -c 0.0001 -analysis U
-libType fr-secondstrand -t paired
-b1 ENCFF780QXW.bam,ENCFF914MST.bam
-b2 ENCFF925GZR.bam,ENCFF842ZDM.bam
-gtf gencode.v19.annotation.gtf
-o <results folder>
```

Custom R scripts were used to select exons from the SE.MATS.Junction-CountOnly.txt file. Enhanced and silenced alternative exons required a p-value < 0.01, an FDR < 0.1 and a $|\Delta \Psi|$ > 0.05 in knockout versus control experiments as in Van Nostrand et al. (2017a). Control exons were selected as those with a p-value > 0.1, an FDR > 0.1, a $|\Delta \Psi|$ < 0.05 and a mean $\Psi$ < 0.9 (the latter to exclude constitutively spliced exons).

### 3.3.4 Generating RNA maps

RNA maps were generated using BEDTools and custom Python and R scripts modified from an pipeline originally written by Dr Nejc Haberman. RNA maps plot normalised crosslink or peak site coverage around regulated exons in a strand-aware manner. BED files of regulated (enhanced and silenced) and control exons are split into two files: one for the 3′ splice site and one for the 5′ splice site. Each of these positions are then extended 300 nt in both directions. The coverage of crosslink or peak sites at each position of these extended 3′ and 5′ splice site BED files is calculated using the BEDTools command:

```
bedtools coverage -sorted -s -b <summed crosslink BED file or peak BED file> -a <extended splice site BED file> -d > <crosslink or peak extended splice site coverage file>.
```
CHAPTER 3. EVALUATION OF METHODS TO STUDY PROTEIN-RNA INTERACTIONS

<table>
<thead>
<tr>
<th>METHOD</th>
<th>REFERENCE</th>
<th>CROSSLINKS</th>
<th>CROSSLINK SITES</th>
<th>RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>iCLIP</td>
<td>Haberman et al. (2017)</td>
<td>45,867,391</td>
<td>27,325,880</td>
<td>1.68</td>
</tr>
<tr>
<td>irCLIP</td>
<td>Zarnegar et al. (2016)</td>
<td>62,811,273</td>
<td>41,389,152</td>
<td>1.52</td>
</tr>
<tr>
<td>eCLIP</td>
<td>Van Nostrand et al. (2017a)</td>
<td>12,197,121</td>
<td>9,186,766</td>
<td>1.33</td>
</tr>
</tbody>
</table>

Table 3.2: Overview of PTBP1 CLIP data sets
Sources and basic sequencing metrics for the three data sets that are compared here. Crosslinks refers to total number of crosslinks after pre-filtered reads that align to rRNA and collapsing PCR duplicates using UMIs. Crosslink sites refers to the number of unique genomic positions at which a crosslink is observed. The ratio is calculated between crosslinks and crosslink sites. Biological replicates have been merged.

Metaprofiles are calculated by summing the depth at each position and normalised by dividing by the number of regulated exons considered. Smoothed RNA maps are generated using a Gaussian smooth with a window of 25, and plotted in the $-300$ to $+50$ positions around the $3'$ splice site and the $-50$ to $+300$ positions around the $5'$ site.

Two metrics are calculated from metaprofiles for CLIP data for silenced exons around the $3'$ splice site to assess the PTBP1 CLIP data quality (Figure 3.3). First, the maximum intronic peak height for the silenced exons is obtained in the $-100$ to $0$ window as a measure of sensitivity. Second, the ratio of the maximum intronic peak height in the $-100$ to $0$ window for the silenced exons to that for the enhanced exons is calculated as a measure of specificity.

3.4 RESULTS

3.4.1 An overview of the data sets

Although the data were produced by different experimental methods, in my computation analysis, I have processed them from the raw sequencing files as uniformly as possible to identify the crosslink positions: in particular with regards to pre-mapping to remove reads that originate from rRNA or tRNA; mapping to the genome and collapsing PCR duplicates using UMIs; and subsequent peak calling approaches.

1Although the irCLIP 5' adapter is modelled on the original iCLIP adapter and so contains both an experimental barcode and a split UMI at the start of the read, this is not in fact used by the
The number of crosslinks detected is approximately 3 to 5 times higher for iCLIP and irCLIP compared to eCLIP (Table 3.2). This is in part a reflection of the sequencing depth of the experiments as PCR duplicates have been collapsed, but the complexity of the cDNA library will also play a role. The ratio of crosslinks to crosslink sites can be used as a general estimate of sensitivity. A lower ratio means that there are fewer pileups of crosslinks, i.e. each crosslink site has only a few crosslinks supporting it, suggesting a data set with sparser, more distributed signal, and so lower sensitivity. Of course the biological pattern of binding will also affect this: for an RBP with a more distributed binding pattern, such as FUS (Rogelj et al., 2012), a low ratio will simply reflect the binding pattern of the RBP. However, we know that PTBP1 binds with peaks of binding, for example upstream of the 3′ splice sites of silenced exons (Attig et al., 2018; Cereda et al., 2014; Haberman et al., 2017; Llorian et al., 2010; Van Nostrand et al., 2017a; Xue et al., 2009), so in this case it is a useful comparator between experiments.

As a preliminary assessment, I considered the regional distribution of the crosslinks. The iCLIP and irCLIP data were from HeLa cells, while the eCLIP data were from HepG2, but the genome-wide distribution of binding should remain broadly similar: the functional relationship between PTBP1 binding and splice sites is unchanged (Chakrabarti et al., 2018), with enrichment of binding in the intronic regions (Attig et al., 2018; Xue et al., 2009).

In all experiments, predominantly intronic binding is seen: 78% for iCLIP, 71% for irCLIP, but only 59% for eCLIP (Figure 3.4). This last compares with 54% in the paired mock eCLIP experiment (where immunoprecipitation for a specific RBP has been omitted). Instead there is an over-representation of exonic binding in eCLIP, with 12% in the CDS. This is much more than iCLIP and irCLIP where the proportion in the CDS is negligible. Indeed, an older study using the original HITS-CLIP method only had 3.5% exonic binding (Xue et al., 2009). There is also an enrichment of exonic binding in the mock, with 7% in the CDS. This suggests there may be contamination. As the eCLIP method neither visualises the protein-RNA bioinformatic pipeline (FAST-iCLIP) to collapse PCR duplicates (Zarnegar et al., 2016). Rather, the first 17 nt of the reads are trimmed and identical sequences collapsed. As discussed previously, this is not an optimal method to collapse PCR duplicates for CLIP data.
Figure 3.4: Regional distribution of crosslinks
The regional distributions of crosslinks across the genome for the three CLIP experiments are shown. As a comparison, the distribution for the size-matched mock eCLIP, where no immunoprecipitation is performed is also shown.

complexes after immunoprecipitation as in iCLIP (Huppertz et al., 2014; König et al., 2010) and irCLIP (Zarnegar et al., 2016), nor performs denaturing purification, as in CRAC (Granneman et al., 2009), one should not assume that the sequenced reads represent only RNAs in contact with the protein of interest: there is a higher risk of non-specific signal. Further examination of the CLIP data after integration with orthogonal function data by means of RNA maps is necessary to understand the quality of the data.

3.4.2 RNA maps to compare PTBP1 CLIP data

As discussed earlier, RNA maps can be used to visualise the positional distribution of assigned CLIP binding sites around regulated landmarks defined by an orthogonal method (Chakrabarti et al., 2018; Witten and Ule, 2011). Hence they are a powerful tool for the assessment of data quality. I have integrated the three PTBP1 CLIP data sets with two data sets that were used to study the role of PTBP1 in alternative splicing regulation.
3.4. RESULTS

1. A high density Affymetrix microarray with probesets corresponding to all well annotated exons and exon-exon junctions (for 2010) that was used in HeLa cells after siRNA knockdown of PTBP1/2 (Llorian et al., 2010). The raw data has subsequently been reprocessed (Cereda et al., 2014; Chakrabarti et al., 2018) and identified 359 enhanced and 419 silenced cassette exons.

2. RNA-seq in K562 cells that was performed after CRISPR knockout of PTBP1 (Van Nostrand et al., 2017a). Using rMATS to assess for alternative splicing events identified 841 enhanced and 776 silenced cassette exons.

3.4.3 Assessing the raw CLIP data quality

For an initial assessment of the data, I examined raw crosslink sites without performing any peak calling (Figure 3.5). The patterns are broadly similar for regulated exons defined by either microarray after siRNA knockdown or RNA-seq after CRISPR knockout. This also supports the earlier statement that the functional relationship between PTBP1 binding and splice sites is similar across cell types. As is evident, iCLIP has the highest sensitivity with 19% of silenced exons containing a crosslink site at the peak position upstream of the 3′ splice site for the microarray-defined exons. This is followed by irCLIP with 11% and eCLIP with 7%. Although the lower sensitivity of eCLIP could, in part, be explained by its having the smallest library size, this does not account for the difference between iCLIP and irCLIP. To assess the specificity of the data, I compared this silenced peak with the same position for the enhanced exons. Here, while we can see that the specificity is again highest for the iCLIP, with a silenced to enhanced ratio of 3.3, it is also reasonable for eCLIP, where the sensitivity was poorest, with a ratio of 2.13. Most notably, irCLIP which had a medium sensitivity has poor specificity indicated by the presence of a similar peak for the enhanced exons as was observed for the silenced, and a ratio of 1.04. The high quality RNA map generated from the raw crosslink sites for iCLIP reflects that in a fully optimised experiment the mapped reads should almost exclusively corre-
Figure 3.5: RNA maps of CLIP crosslinks around regulated exons as identified by microarray (A, C, E) and RNA-seq (B, D, F). The RNA maps of iCLIP, irCLIP, and eCLIP raw crosslink sites around regulated exons are annotated as described in Figure 3.3.
3.4. RESULTS

...spond to the sites of RBP-RNA interactions, with minimal noise from non-specific background (Chakrabarti et al., 2018). This is due to the many experimental steps targeted to remove noise from the sequencing library (Lee and Ule, 2018).

The regional differences discussed in the previous section are also seen here, with a rise in signal for eCLIP within the exonic regions for all categories of regulated exon. This is more evident in the RNA-seq than the microarray RNA maps, possibly a reflection of the higher number of exons identified in the latter, but could also reflect a higher false positive rate in defining alternatively spliced exons with the transcriptome wide analysis. In contrast, particularly for iCLIP, there is minimal signal in the corresponding regions for both sets of RNA maps.

3.4.4 Assessing the effect of peak calling on data quality

Next, I assessed to what extent peak calling could improve the data quality, using the iCount peak caller with the same parameters for all three data sets to ensure a fair comparison (Figure 3.6). Here, there was only a small change in the iCLIP RNA map from that using the raw crosslink position (Figure 3.5) with a slight increase in sensitivity to 23% from 19% for the microarray RNA map and a slightly larger increase in the specificity with a ratio of 5.16 from 3.3. Again this reiterates the minimal background signal in this data set. For eCLIP, peak calling had a noticeable impact on the specificity with a marked improvement to a ratio of 4.95 from 2.13 for the microarray RNA map, albeit without any improvement in sensitivity. While the improvement in eCLIP specificity was also noted for the RNA-seq RNA map, this was at a cost to sensitivity which had fallen to 4% from 6%. Interestingly, peak calling also reduced the exonic signal in both eCLIP RNA maps, more so for the microarray, although this remained higher than for iCLIP. This supports the hypothesis that the exonic signal is indicative of background noise. Finally, for irCLIP there was a very small improvement in specificity to 1.12 from 1.04 for the microarray RNA map at a cost of sensitivity which fell to 7% from 11%.
Figure 3.6: RNA maps of CLIP peaks around regulated exons

RNA maps of iCLIP, irCLIP and eCLIP peak sites around regulated exons as identified by microarray (A, C, E) and RNA-seq (B, D, F). Peak calling was performed with iCount using a 15 nt window. The RNA maps are annotated as described in Figure 3.3.
3.4.5 Assessing the effect of different peak callers on data quality

In order to understand the effect of different peak callers on the sensitivity and specificity of the data, I compared three widely used CLIP peak callers: iCount (Curk et al., 2017), Piranha (Uren et al., 2012) and CLIPper (Lovci et al., 2013) and one clustering algorithm originally developed for CAGE (cap analysis of gene expression) data: Paraclu (Frith et al., 2008). I used the iCLIP data as it was of the highest quality\textsuperscript{2}. Where possible a range of parameters were trialled for each tool to identify optimal settings, and representative RNA maps using the microarray-defined exons are shown in Figure 3.7. Here, we can see that all peak callers fulfil their primary aim of increasing data specificity. Piranha has the greatest effect with a three-fold increase in the specificity ratio, followed by Paraclu. CLIPper and iCount show the smallest effect, although this is still a noticeable improvement. However, these improvements in specificity need to be balanced against the effects on sensitivity. As expected, the ranking of tools here is reversed: with iCount having the highest sensitivity with 34% of exons and Piranha the least with only 9% of exons. Hence the dramatic improvement in specificity with Piranha comes with a major undesired effect on sensitivity, with a nearly 3-fold fall compared to the raw crosslink sites. The optimal compromises are probably achieved for Paraclu, with a 1.7-fold increase in specificity and 1.4-fold in sensitivity; very closely followed by iCount, with a 1.2-fold increase in specificity and 1.8-fold in sensitivity.

3.4.6 Assessing optimisation of peak calling parameters

Next, I explored the optimisation of the two best-performing peak callers, Paraclu and iCount, in more detail. There were two angles, I considered: first, how different settings affected the RNA map profiles and how much ‘tuning’ the tools required to obtain optimal settings; and second, whether there was a difference with CLIP data

\textsuperscript{2}Although the run times were not formally assessed, for the tested data there was considerable variability. Using 1 processor, Piranha and Paraclu took 1 to 2 minutes and iCount took approximately 7 hours; using 20 processors CLIPper took nearly 7 days, largely on account of the spline fitting algorithm.
Figure 3.7: RNA maps showing the effect of different peak callers.
RNA maps of iCLIP peak sites as called by different peak callers around regulated exons as identified by microarray. They are ordered with raw crosslink sites at the top, and then by decreasing sensitivity. The relevant parameters for each peak caller is indicated in the plot title. The RNA maps are annotated as described in Figure 3.3.
3.4. RESULTS

sets of different qualities. I again used the microarray-derived exons and compared
the iCLIP and eCLIP data after peak calling with systematic step-wise increases in
window size for iCount and maximal cluster size for Paraclu.

With iCount (Figure 3.8), peak calling always improved upon the raw signal,
both in terms of sensitivity and specificity. Indeed, good RNA maps were obtained
at all tested windows both with iCLIP and with eCLIP data. However, the maximal
fold-improvements from the raw data were more pronounced for eCLIP than for
iCLIP, both for sensitivity (2.7 versus 2.1, respectively) and for specificity (2.3 versus
1.6, respectively). For each window size, there was a greater increase in fold-
change for specificity for eCLIP compared with the same window size for iCLIP.
This likely reflects the higher data quality of the iCLIP, with overall less scope for
improvement. On account of this, it is possible to select a 15 or 20 nt optimal window
and compromise more on the specificity fold-change to achieve a greater sensitivity
for iCLIP (15 nt: 1.2 and 1.8 fold-change for specificity and sensitivity, respectively;
20 nt: 1.1 and 1.9 fold-change for specificity and sensitivity, respectively), whereas
for eCLIP there is a more narrow range of optimal window at 15 nt (1.6 and 1.9
fold-change for specificity and sensitivity, respectively) with a greater drop-off with
smaller or larger windows.

There was a different picture with Paraclu (Figure 3.9). At the smaller maximal
cluster sizes (10 to 50 nt), there was in fact a marked reduction in sensitivity for both
iCLIP and eCLIP. Although this was usually concomitant with an increase in speci-
ficity, for iCLIP using a 10 nt setting in fact also decreased specificity compared to
the raw data. Interestingly, although reciprocal changes in sensitivity and specificity
were observed going from 50 to 100 nt, a further increase to 200 nt resulted in an
increase in specificity both for iCLIP and eCLIP. Inappropriate parameters have a
much greater effect on the outcome for Paraclu compared with iCount. However,
ultimately the optimal settings were the same for both iCLIP and eCLIP with a 200
nt maximal cluster size.

The simplicity of identifying optimal parameters and robustness of a tool to a
range of settings is not inconsequential for CLIP analysis. Here, I have assessed only
one RBP, PTBP1, with its particular pattern of binding and binding site sizes. Other
Figure 3.8: RNA maps showing the effect of different iCount peak calling parameters
RNA maps of iCLIP (left panel) and eCLIP (right panel) peak sites around regulated exons as identified by microarray. From top to bottom there are increasing window sizes for iCount peak calling as indicated in the titles. The RNA maps are annotated as described in Figure 3.3.
Figure 3.9: RNA maps showing the effect of different Paraclu peak calling parameters. RNA maps of iCLIP (left panel) and eCLIP (right panel) peak sites around regulated exons as identified by microarray. From top to bottom there are increasing maximum cluster sizes for Paraclu peak calling. The RNA maps are annotated as described in Figure 3.3.
RBPs will have different modes: either with narrower sites, such as PRPF8 (Blazquez et al., 2018), or wider sites, such as FUS (Rogelj et al., 2012). Hence, I speculate that settings optimised for Paraclu for PTBP1 may be less transferrable to CLIP data from other RBPs as compared with iCount, which appears more forgiving.

3.5 CONCLUSIONS

This chapter highlights the intimate relationship between experimental and computational halves of methods developed to analyse protein-RNA interactions. First, I explored the data quality of different CLIP methods and the impact of different tools and parameters both on data quality, but also how they might affect biological interpretation of functional binding patterns through RNA maps. Here, it was remarkable that for a high quality dataset, such as that obtained for PTBP1 using iCLIP, robust biological conclusions could be drawn even without peak calling. While peak calling can improve somewhat the specificity of data sets such as that obtained by eCLIP, experimental visualisation of the protein-RNA complexes enables quality control and the best optimisation of the experimental set up. Background correction with a size-matched mock IP sample for the purpose of improving the specificity is only necessary where this step has been omitted, but is not a substitute for a high quality experiment. Finally, different peak calling tools have a range of effects on the RNA maps produced from the same data. The best compromises between sensitivity and specificity were seen with Paraclu and iCount. Most peak callers require user-defined parameters that should be defined on the basis of knowledge of the binding pattern of the RBP under study. Of these two best performing peak callers on this data set, iCount is more lenient in this regard than Paraclu, although the latter may be slightly more uniform across data of different qualities. A through assessment of data quality is important before interpreting results from CLIP data and RNA maps are a powerful tool to do so.
# Development of Methods to Study Protein-RNA Duplex Interactions

1. **Summary of the chapter** ........................................... 134
2. The hiCLIP method to study protein-RNA duplex interactions .............. 134
   2.1 The innovations of the hiCLIP method .................................. 135
   2.2 Hypotheses of the sources of ‘lost’ duplexes during data analysis ...... 137
3. **Methods** .............................................................. 141
   3.1 Processing of hiCLIP data ............................................. 141
   3.2 Alignment of non-hybrid reads ...................................... 141
   3.3 Identification of putative duplexes from aligned non-hybrid reads ...... 142
   3.4 Identification of putative duplexes protected from RNase digestion ... 143
   3.5 Identification of confident duplexes ................................ 144
4. Development of improved computational methods for hiCLIP data analysis .... 144
   4.1 The linker adapter may be partially degraded .......................... 144
   4.2 The linker adapter is not necessary to form hybrid reads .............. 144
   4.3 Short-range duplexes are protected from RNase digestion ............. 147
   4.4 Extending the yield of confident hiCLIP duplexes ..................... 150
5. **Conclusions** .......................................................... 151
4.1 SUMMARY OF THE CHAPTER

This chapter explores the analysis of genomic data from the hiCLIP (hybrid iCLIP) experimental method that has been used to study interactions between double-stranded RNA binding proteins (dsRBPs) and RNA duplexes or structured RNA (Sugimoto et al., 2015). I start by introducing the experimental modifications that were incorporated into the iCLIP method to enable the detection of RBP-bound RNA duplexes. However, the method is limited by low sensitivity, with a poor recovery of hybrid reads from the sequencing library. I re-assess the original experimental method and identify three possible experimental inefficiencies that result in the mis-calling of hybrid reads as non-hybrid by the published analysis pipeline (Sugimoto et al., 2017). Implementing new computational methods to tackle these issues recovers these ‘lost’ hybrids, ultimately leading to a 7-fold increase in the number of confident duplexes from the same original data. In addition to ensuring that the maximum is obtained from the experimental data, this is also necessary to undertake integrative genomic studies with other transcriptomic data sets to derive generalisable features and patterns as in Chapter 5.

4.2 THE HICLIP METHOD TO STUDY PROTEIN-RNA DUPLEX INTERACTIONS

The hiCLIP method was developed by incorporating proximity ligation into the iCLIP method (discussed in more detail in Chapter 3) to identify RNA duplexes bound by an RBP of interest (Sugimoto et al., 2017, 2015). Here, I focus on the details of the experimental method (Figure 4.1); hiCLIP is contextualised within the framework of current genomic methods to study RNA structure in Chapter 5.
4.2. THE HICLIP METHOD TO STUDY PROTEIN-RNA DUPLEX INTERACTIONS

4.2.1 The innovations of the hiCLIP method

The extraction of RBP-bound RNA fragments follows the same principles as detailed for CLIP in Chapter 3 and is detailed in Figure 4.1 (steps 1–2): with UV-C crosslinking of protein-RNA complexes, cell lysis, partial RNA digestion and immunoprecipitation of the complexes. In the original hiCLIP method, a relatively mild cell lysis buffer was used to capture the cytoplasmic compartment and to ensure that the buffer did not disrupt the duplexes before proximity ligation of the two arms. RNase I preferentially digests single stranded RNA, hence is used to preserve the duplexes.

Next, the first adapter ligation is performed (step 3). The key innovation of hiCLIP was the use of two different adapters: a second linker adapter was introduced in addition to the usual iCLIP SeqRv adapter (Lee and Ule, 2018; Sugimoto et al., 2015) – adapters B and A respectively in Figure 4.1. This is a flexible 19 nt ribonucleotide adapter with a reversible 3’ phosphorylation block\(^1\). The use of equal concentrations of each adapter results in a 50% probability of a duplex having a

---

\(^1\)In contrast, the SeqRv adapter has an irreversible 3’ block to prevent the addition of additional molecules (Huppertz et al., 2014; Lee and Ule, 2018; Sugimoto et al., 2017)
SeqRv adapter ligated to one arm, and the linker adapter to the other. The 3′ block is then removed from the linker adapter (step 4) and a second round of ligation performed (step 5). This is the proximity ligation step: the linker adapter now has a free 3′ end to enable joining of the two duplex arms into one molecule.

These additional steps were introduced for three reasons (Sugimoto et al., 2017). First, a flexible region of single RNA next to the RNA duplex is likely to be necessary to enable proximity ligation because of physical constraints, with evidence that at least eight nucleotides in an RNA are needed for efficient circularisation (Kaufmann et al., 1974; Sugimoto et al., 2015). As such an overhang may not exist after partial RNA digestion down to the protected RNA duplex; the additional flexible linker adapter can fulfil this role to enable efficient proximity ligation. Second, the use of the linker ensures that all hybrid reads have arisen as a result of the controlled ligation reaction. Otherwise, potential endogenous ligase activity could result in hybrids that do not represent RNA duplexes (Cimino et al., 1985). Third, identification of the two arms of the duplex using direct proximity ligation is computationally challenging (Travis et al., 2014) and there is the risk of erroneous assignment of duplex arms. The use of the linker ensures unambiguous delineation of the two arms in the hybrid read before alignment using the hiclipr bioinformatic pipeline (Sugimoto et al., 2017).

A priori information about the nature of the bound duplexes (e.g. AGO1 binding miRNA-mRNA duplexes (Helwak et al., 2013)) is therefore not necessary to identify dsRBP-bound RNA duplexes – an agnostic approach can be used.

After adapter ligation, the subsequent experimental steps (step 6) proceed as for iCLIP (Lee and Ule, 2018) to generate the cDNA library for high-throughput sequencing. Bioinformatic analysis (step 7) uses the linker adapter sequence to separate out the two duplex arms unambiguously, which can then be independently mapped to define the positions of each duplex arm on RNA transcripts (Sugimoto et al., 2017).
4.2. THE HICLIP METHOD TO STUDY PROTEIN-RNA DUPLEX INTERACTIONS

Figure 4.2: Sources of ‘lost’ duplexes during data analysis
The three hypotheses of how ‘lost’ duplexes may have arisen are shown. On the left are the expected experimental outcomes for the relevant steps (as in Figure 4.1) and on the right in the shaded box is the corresponding alternative outcome that would not have been detected by the existing analysis approach: (A) the linker adapter (adapter B) is partially degraded; (B) the linker adapter fails to ligate, but proximity ligation still occurs between the two duplex arms (dashed line); (C) the duplex loop is protected by the protein and not digested by the RNase.

4.2.2 Hypotheses of the sources of ‘lost’ duplexes during data analysis

The original hiCLIP data (Sugimoto et al., 2015), reprocessed using my hiclipr R package (Sugimoto et al., 2017), identified 5,833 unique non-redundant duplexes. The reprocessed results are essentially almost identical to those in the original publication (Sugimoto et al., 2017). However, only 774 of these 5,833 were supported by more than 1 hybrid read and deemed confident duplexes. The paucity of robust duplexes is a limitation for integrative genome-wide analyses. I reviewed the data analysis, with careful consideration of the experimental method, to maximise the yield of confident duplexes.

One of the first steps of the original computational method was to categorise the sequenced reads into hybrid and non-hybrid reads, based on the presence of
a linker adapter within the read. However, hybrid reads only comprised 1-2% of the whole sequencing library (Sugimoto et al., 2015). I aimed to show that true biological duplexes were in the sequenced reads, but were erroneously categorised as non-hybrid. I generated three hypotheses with reference to the experimental steps (Figure 4.2) as to how this might have occurred and tested each of them in turn. They are listed in order of where in the computational analysis they are encountered (which incidentally is the reverse of the experimental steps from which they originate).

1. In step 4 (Figure 4.2A), the reversible 3′ phosphorylation block of the linker adapter (adapter B) is removed prior to the second ligation reaction. The 3′ phosphorylation of the linker adapter was, however, imperfect leading to 20–30% of reads containing the linker adapter directly followed by the SeqRv adapter in place of the SeqRv adapter alone (Sugimoto et al., 2015). Moreover, the adapter (is comprised of ribonucleotides and so is more susceptible to degradation than a DNA linker. This raises the possibility that a few nucleotides may have been degraded from the end of linker adapter and these hybrids were missed in the read categorisation.

2. In step 3 (Figure 4.2B), the SeqRv and linker adapters are ligated. During ongoing development of the experimental methods, it was noted that the efficiency of ligation of the linker adapter to one of the free duplex ends was poor (Dr Cristina Milliti, personal communication). However, we know from subsequently developed proximity ligation approaches that, depending on the level of RNase digestion and hence level of single-stranded RNA remaining, a linker may not be necessary for the intermolecular ligation described in step 5 (Aw et al., 2016; Lu et al., 2016; Sharma et al., 2016). I hypothesised that there were a proportion of duplexes where only the SeqRv adapter was ligated with a failure of ligation of linker adapter at step 2. However, these duplexes were still subjected to the ligation reaction mix in step 5 and so were able to undergo proximity ligation of the two fragments even in the absence of a linker adapter. These reads would then
be grouped with the non-hybrid reads.

3. In step 2 (Figure 4.2C), RNase is used to partially digest the RNA to separate the free ends of the two duplex arms. The unprotected, primarily single-stranded loop of RNA between the two duplex arms is removed, leaving behind the duplex that is protected by the crosslinked RBP in two fragments. However, I hypothesise that in the case of short loops, they are sterically protected from the action of the RNase by either the RBP or the RBP-associated protein complex and thus remain undigested. Another possibility is that a short loop may be more difficult to digest given the proximity of the base-paired stem. Consequently, there is no longer the possibility for the ligation of two adapters, one to each fragment as in step 3, and only one adapter can be ligated instead. Although both SeqRv and linker adapters will ligate with equal probability, only the SeqRv adapter contains the priming site for the reverse transcription (step 6) and so only these will contribute to the sequencing library. However, they will be mis-assigned as non-hybrid reads.

To summarise, I predict there will be four different types of sequenced read: i) hybrid with a linker adapter (full-length or shortened), containing a duplex; ii) hybrid without a linker adapter, containing a duplex; iii) non-hybrid, containing a duplex adjacent to the crosslink position; and iv) non-hybrid, not containing a duplex. In evaluating these hypotheses, I aimed to develop computational methods to recover these three categories of ‘lost’ duplexes to augment the yield of the hiCLIP data. This is necessary to overcome the existing limitations of the data and yield sufficient numbers of confident duplexes to permit further integrative analysis with complementary and functional data to explore the biological roles of these structures as exemplified later in Chapter 5.
Pre-processed sequenced hiCLIP reads with SeqRv adapter removed

Linker adapter present? (4.3.1)

Yes

No

Degraded linker adapter present? (4.3.1)

Yes

No

Proximity ligation without linker adapter? (4.3.3)

Yes

No

Duplex protected from RNase digestion? (4.3.4)

Yes

Hybrid reads (4.3.5)

No

Confident duplexes

Non-hybrid read

---

Figure 4.3: The computational analysis workflow
The processing stages that the hiCLIP sequenced reads undergo to identify hybrids highlighting the decision points. In red are the corresponding subsections in the methods.
4.3 METHODS

An overview of the stages comprising the analysis workflow are shown in Figure 4.3 and further expanded upon below.

4.3.1 Processing of hiCLIP data

The basic processing steps of adapter removal, collapsing of PCR duplicates with UMIs, the mapping of hybrids with the linker adapter and the identification of confident duplexes were performed using the R package hiclipr, which I developed from the original analysis approaches in (Sugimoto et al., 2015) and as detailed in Sugimoto et al. (2017). Here, I focus on the methods specific to the new hiCLIP analysis approaches presented in this chapter.

4.3.2 Alignment of non-hybrid reads

STAR (Dobin et al., 2013) was used to align non-hybrid reads, in a four-stage alignment strategy similar to the hybrid reads. Custom STAR indices matching those generated for Bowtie in hiclipr were created using the same FASTA files, taking care to scale down the parameter \(-\text{genomeSAindexNbases}\) for the smaller indices. First, reads were aligned to a custom rRNA and tRNA index. Second, unmapped reads were then aligned to a custom mtDNA index. Third, unmapped reads were aligned to a custom transcriptome. This was created from the GRCh37 Ensembl 75 annotation, and contained one representative transcript per gene. For protein-coding genes, the longest protein coding transcript was selected with ties broken with the most exons and then the longest 3' UTR. For non-coding genes, the longest non-coding isoform was selected. Fourth, unmapped reads were aligned to the GRCh37 genome with the Ensembl 75 annotation. For all stages the same STAR parameters as in Lu et al. (2016) were used:

```
--outReadsUnmapped Fastx --outFilterMultimapNmax 100 \ 
```
For each stage, this outputs two BAM files: one for regularly aligned reads and one for chimeric aligned reads.

### 4.3.3 Identification of putative duplexes from aligned non-hybrid reads

The identification of putative duplexes was focused on the reads that aligned to the transcriptome and performed using custom functions written in R. First, it is important to note that STAR also outputs chimeric aligned reads as unmapped reads, so any reads that aligned to the transcriptome, but had also been aligned as a chimeric read in the previous phases were filtered. Then valid putative duplexes were selected on the basis that the read mapped in two parts in the forward orientation to the transcriptome. For chimeric reads, further checks were performed to exclude reads where different parts of the read aligned to the same genomic location. Finally, reads that had the same UMI and left and right arm start positions were collapsed as PCR duplicates.

RNAduplex from the ViennaRNA package (Lorenz et al., 2011) was used to anneal the two duplex arms with the command: 

```
RNAduplex -noLP <left arm sequence>
    <right arm sequence> > <dot-bracket output>
```

and predict the duplex structure. A custom R function was then used to parse the dot-bracket output to identify which positions in each of the duplex arm sequences corresponded to the first and last paired position in the predicted duplex. The duplex arm coordinates were trimmed according to these positions to identify the duplex region within the arm. Of note this is different to the original method where the longest continuous paired stretch of positions within the sequence was considered the duplex (Sugimoto et al., 2017, 2015).

RNAduplex was re-run on the trimmed duplex arms to calculate the minimum free energy of hybridisation using the same command as above. Shuffled sequences
with preserved dinucleotide frequency were obtained using uShuffle (Jiang et al., 2008) and the command `uShuffle -seed 42 -k 2 -n 1 -s <duplex arm sequence>`. A Gaussian mixture model with three components was fitted to the minimum free energy distribution with an expectation maximisation algorithm using the mixtools R package (Benaglia et al., 2009) and the `normalmixEM` function. Starting values of $\mu$ and $\sigma$ were selected using the default methods. This was necessary to differentiate true duplexes from false positives as discussed later.

4.3.4 Identification of putative duplexes protected from RNase digestion

Non-hybrid reads, excluding those which no-linker duplexes had been detected in the previous step, were used for this analysis. First, alignments were converted from transcriptomic to genomic coordinates using the hiclipr package (Sugimoto et al., 2017). This was necessary to perform peak calling using iCount to identify significant peaks of binding. Summed crosslink BED files were generated from these coordinates and iCount run with a 10 nt half-window setting as above. Identified peaks were then extended to a uniform 100 nt, while keeping the start position (the crosslink site) fixed. FASTA files were generated containing the sequences for each of these regions.

LocalFold (Lange et al., 2012) was used to calculate the base-pairing probability matrix. This is a modification of the RNAplfold algorithm from the ViennaRNA package (Bernhart et al., 2006; Lorenz et al., 2011). The FASTA file was used as input to generate pairing dotplot PostScript files for each sequence using the command: `localfold.pl -noacc -seqfile=<FASTA>`. From each PostScript file, the value of the width of the squares above the diagonal was extracted using AWK. A custom R script was run to square these values (which gave the base-pairing probability for that pair of positions) and convert the data frame of values into a 100 $\times$ 100 matrix. These matrices were filtered using a custom R function that identified duplexes in the matrix on the basis of a minimum pairing probability threshold, a maximum distance from position 1 (the crosslink) and the minimum continuous
stretch of paired positions necessary to call a duplex. Finally, these duplexes were converted back to transcriptomic coordinates.

4.3.5 Identification of confident duplexes

Confident duplexes, i.e. duplexes with more than one hybrid supporting them (Sugimoto et al., 2017, 2015), were selected using an updated version of the FindIslands function from the hiclipr package that enabled parallelisation to permit a reasonable runtime.

4.4 Development of improved computational methods for Hi-CLIP data analysis

4.4.1 The linker adapter may be partially degraded

Hybrid reads are identified by searching for the full length linker adapter (with one permitted mismatch, but no deletions). Therefore, reads with shorter sequences of a partially degraded adapter would have been missed and as a result labelled as non-hybrid reads. Examining the sequencing libraries for a shortened variant of the linker adapter recovered a further 8,596 hybrid reads (14% of the original total) for alignment and PCR duplicate removal.

4.4.2 The linker adapter is not necessary to form hybrid reads

Next, I assessed whether proximity ligation had occurred in the absence of linker adapter ligation: either due to an overestimation of the inefficiency of circularisation (Kaufmann et al., 1974; Sugimoto et al., 2015) or because a long-enough single stranded stretch of RNA (i.e. of the loop) remains after partial RNase digestion to form the bridge between the two arms.

One of the other reasons for including the linker adapter was to enable easy
deconvolution of the two duplex arms from a read as the known adapter sequence would be identifiable between the sequence for either arm. Detecting the join between the two arms without this marker is more challenging. As the ligation of the SeqRv adapter could occur on the 3′ of either duplex arm fragment, there are two possibilities as to the orientation of the duplex arms in the sequenced read that should in theory occur at equal probabilities (Figure 4.4). The first possibility is that the proximal arm (5′ in genomic space) is at the beginning of the read and then followed by the distal arm (3′ in genomic space); in other words the orientation of the arms in the read matches genomic orientation. The second possibility is that the proximal arm is at the end of the read and preceded by the distal arm; in other words the orientation of the arms in the read are in a reverse orientation.

To identify both these sets of duplexes, I used the STAR aligner (Dobin et al., 2013) which had been previously been shown, with certain settings, to be able to identify hybrid reads generated by proximity ligation methods (Lu et al., 2016). This strategy leveraged two abilities of STAR that were able to account for each of the two read orientation possibilities: first, the ability to align to non-canonical splice junctions and second, the ability to map chimeric transcripts. Essentially the read is treated somewhat like a transcript with a non-canonical splice junction for genomic
CHAPTER 4. DEVELOPMENT OF METHODS TO STUDY PROTEIN-RNA DUPLEX INTERACTIONS

orientation hybrid reads, and like a chimeric or fusion transcript for the reverse orientation hybrid reads. Hence both types of reads are able to be mapped by STAR, and from this output duplex arms from either orientation read can be deconvolved.

I developed a multi-stage non-hybrid read mapping pipeline, along the same lines as the original one for hybrid reads that I implemented in the hiclipr package (Sugimoto et al., 2017, 2015). This first filtered hybrids originating from rRNA or tRNA, and then identified hybrids originating from the transcriptome (rather than the genome to avoid mis-assignment of reads spanning true splice junctions as hybrid reads). With this approach, after collapsing PCR duplicates, I recovered 41,562 putative transcriptomic duplexes for the high RNase condition and 24,383 for the low RNase condition. By comparison, there were 6,004 and 2,295 for the two RNase conditions respectively obtained from the hybrid reads.

There are two potential sources for false positives within these putative duplexes. First, RNA hybrids may have been created as a result of endogenous ligase acting on proximate RNA fragments during the early steps of the protocol. This was another reason for the addition of the linker adapter in the original hiCLIP method. This possibility is further compounded as random ligations could also have occurred when the RNA fragment from which these reads originate were subjected to the controlled second ligation reaction. Second, given the need to relax the STAR settings to detect non-canonical splice junctions, there is the possibility of erroneously mapped reads confounding the results.

To assess for both these possibilities, I assessed the thermodynamic stability of the putative duplexes (Figure 4.5). Putative duplexes from both RNase conditions were combined for this quality control assessment. I calculated the minimum free energy of hybridisation of the two duplex arms computationally based on their sequences. This was compared with control sequences obtained by shuffling the sequences of each of the arms while preserving dinucleotide content. Maintaining the dinucleotide frequency is important when defining a background signal to assess thermodynamic stability of RNA secondary structures (Babak et al., 2007). The density distribution of binding energies was bimodal (Figure 4.5B), with one peak at a similar binding energy to that of the duplexes identified from the hybrid reads.
4.4. Development of improved computational methods for HICLIP data analysis

reads (Figure 4.5A), and the other peak at a similar binding energy to the shuffled control. This suggests that some of these putative duplexes do reflect false positives either from mapping or ligation artefacts. However, the majority of duplexes are within the low binding energy sub-population and thus a large number of duplexes stand to be recovered from the non-hybrid reads.

To categorise the putative duplexes into these two sub-populations, I fitted a three-component Gaussian mixture model to the distribution of binding energies (Figure 4.5D). (The third component was necessary to account for the low energy tail of the distribution.) The distribution corresponding to the high binding energy duplexes has a mean energy of \(-6.7\) kcal/mol, whereas the two distributions corresponding to the low and very low binding energy duplexes have mean energies of \(-20.9\) kcal/mol and \(-36.8\) kcal/mol respectively. The weight (\(\lambda\)) of the high binding energy distribution in the mixture model is 0.24, which can be interpreted as meaning that nearly 75% of the duplexes potentially being biological. However, I use a stringent threshold to filter duplexes that have a greater than 5% posterior probability of belonging to the high binding energy distribution to minimise the number of false positive duplexes taken forward. This results in 39,449 out of the original 65,945 duplexes (59.8%) remaining, with the binding energy distribution now similar to that for the duplexes from the hybrid reads (Figure 4.5A, C).

4.4.3 Short-range duplexes are protected from RNase digestion

To explore the possibility of short-range duplexes, I examined the non-hybrid reads remaining after the previous interventions. The reads containing these duplexes are truly non-hybrid reads, as the two duplex arms have not been separated and so the read will fully align with the genomic sequence uninterrupted. Hence they would have not been detected in the preceding analysis. A different approach was needed.

The non-hybrid reads are in essence very similar to iCLIP reads. I therefore identified the crosslink positions and peak called using iCount (Curk et al., 2017) using the same approach as described in Chapter 3. This identified a set of significant
Figure 4.5: Binding energy of duplexes identified with and without a linker adapter. Distributions of computationally predicted binding energies for hybridisation of the duplex arms (green) are compared with shuffled controls that preserve dinucleotide frequency (grey) are shown for duplexes obtained from reads containing a linker adapter (A) from reads without a linker adapter (B) and after filtering duplexes obtained from reads without a linker adapter for false positive duplexes (C) using a three component Gaussian mixture model (D) that fit three distributions: light blue: \( \mu = -36.8, \lambda = 0.08 \); dark blue: \( \mu = -20.9, \lambda = 0.67 \); red: \( \mu = -6.65, \lambda = 0.24 \). For the mixture model, the red distribution overlaps the shuffled control and reflects false positive duplexes.

crosslink sites around which a putative short range structure could occur. Given that we know that crosslinking occurs at single stranded bases, and assuming a short stem-loop structure, there are three possibilities as to where the crosslink might occur in relation to a duplex: i) in the region immediately 5' to the duplex (i.e upstream); ii) in the loop; and iii) in the region immediately 3' to the duplex (i.e. downstream).

I focused on exploring the possibility of the first option, as this would be more straightforward and less open to ambiguous results as for the other two options where a duplex arm located 5' to the crosslink position could represent either proximal or distal arms, i.e. the second or third options. I used LocalFold (Lange et al., 2012) to predict the base pairing probabilities of the 100 nt sequence downstream of each crosslink position. This approach calculates minimum thermodynamic free energy structures to derive these probabilities. For each sequence, this gave a pairing
Figure 4.6: Computational prediction of short-range duplexes in non-hybrid reads

Base pairing probability matrices of the 100 nt downstream of binding peaks are visualised as heatmaps. The same 100 nt sequence is shown on both axes, hence the heatmaps would be symmetrical around the diagonal (white line), but the half below the diagonal has been removed for visual clarity. Each square in the heatmap is coloured according to the probability of those two positions being paired. Examples are given (white arrowheads) of a short 10 nt stem loop structure starting at the 2nd nt after the crosslink (A) and a 12 nt stem loop with two small symmetrical internal loops starting at the 3rd nt after the crosslink (B). An average heatmap of all the sequences is shown before (C) and after (D) selection of the short-range duplexes in non-hybrid reads, demonstrating the enrichment of short stem-loop structures after filtering all candidate crosslink sites.
probability matrix, from which it was possible to detect high probability stem loop structures (Figures 4.6A, B).

By parsing all of these matrices, I identified 1,284 putative duplexes from 83,088 significant crosslink positions for the high RNase condition and 1,140 putative duplexes from 7,079 significant crosslink positions for the low RNase condition. These stem loop structures were selected with requirements for: i) a minimum pairing probability for each position of \( \geq 0.9 \); ii) a minimum stem length of 8 nt; and iii) the stem to start within 10 nt of the crosslink position (to correspond with the iCount peak calling window). Overall, putative duplexes were recovered from 16% of crosslink positions (Figures 4.6C, D).

4.4.4 Extending the yield of confident hiCLIP duplexes

Finally, I combined all of these newly identified putative duplexes with those obtained from the hybrid reads. I focused the remaining analysis on intragenic duplexes, that is duplexes where both arms are from the same RNA: this comprised 92% of the duplexes obtained in Section 4.4.2, and by definition all of those obtained in Section 4.4.3. I then used these to identify a set of non-redundant hiCLIP duplexes that were supported by more than one duplex-containing read. This resulted in a final set of 5,419 confident duplexes, a 7-fold increase over the original 774, for which confident duplexes were defined in the same way (Figure 4.7). Interestingly, 60% of these had not been identified in any form in the original characterisation. This may be because they were false positives, although the use of the linker should reduce the probability of random or spurious ligations. An alternative explanation is that they might be duplexes from less abundant transcripts, hence they are difficult to detect with CLIP methods and so under-represented in our data set.
4.5. CONCLUSIONS

This chapter reiterates the intimate relationship between experimental and computational halves of methods developed to analyse protein-RNA interactions. I identified limitations in the hiCLIP method and developed computational solutions to overcome them, largely centred on the idea that the linker adapter was not ligated for all the duplexes in the sequencing library. First, I found that proximity ligation without a linker adapter can and does occur at a higher frequency than with a linker adapter and moreover, valid duplexes can be identified from hybrid reads without the use of the linker adapter to deconvolve the two duplex arms. Second, I identified short-range stem-loop structures bound by the RBP using computational prediction of structures around crosslink sites that were likely protected from RNase digestion to separate the two duplex arms and thus previously undetected. Application of these additional computational methods results in an overall 7-fold increase in the yield of confident RBP-bound duplexes from the same original hiCLIP data set.

Here, I have discussed the hiCLIP method and results without reference to the RBP under investigation. In the next chapter, I use the integrative analyses that this expanded set of duplexes now affords, to offer some biological insight.
INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFOEN PROTEINS

5.1 Summary of the chapter ............................................. 154
5.2 Introduction ....................................................... 155
  5.2.1 The Staufen family of RNA binding proteins .............. 155
  5.2.2 Staufen functions: localisation, stability, translation .... 158
  5.2.3 Experimental methods to assess RNA structure ........... 162
5.3 Methods: STAU1 .................................................. 167
  5.3.1 Processing of PARIS data .................................... 167
5.4 Results: STAU1 .................................................... 168
  5.4.1 Contextualising STAU1-bound duplexes .................... 169
  5.4.2 STAU1 binding and RNA metabolism ....................... 173
5.5 Discussion: STAU1 ............................................... 177
  5.5.1 Differences between hiCLIP and PARIS are a combination of biological and technical effects .................................................. 177
  5.5.2 STAU1-bound structures may have roles in RNA metabolism ....................... 178
5.6 Methods: STAU2 .................................................. 179
  5.6.1 The hiCLIP experimental method adapted to tissue .......... 179
  5.6.2 hiCLIP data pre-processing and read alignment ........... 181
  5.6.3 Hybrid read identification and duplex delineation ........ 182
  5.6.4 Gene ontology analysis ..................................... 183
  5.6.5 Processing of rat brain RNA-seq data ....................... 183
  5.6.6 Miscellaneous analyses .................................... 185
5.7 Results: STAU2 ................................................. 185
  5.7.1 Extension of the Rattus norvegicus annotation .............. 186
  5.7.2 Characterising STAU2 binding in rat brain ................. 188
  5.7.3 hiCLIP identifies STAU2-bound duplexes in rat brain .... 192
  5.7.4 The validity of the detected STAU2-bound duplexes ........ 197
  5.7.5 The dynamics of STAU2-bound duplexes .................... 199
  5.7.6 Structures of STAU2-bound duplexes ....................... 203
  5.7.7 Interactions with other RBPs ................................ 205
5.8 Discussion: STAU2 ............................................... 207
  5.8.1 hiCLIP detects valid RNA duplexes in brain tissue ......... 207
  5.8.2 STAU2 binding to RNA duplexes changes across development ... 211
  5.8.3 Interactions with HuR and Rbfox1 .......................... 213
  5.8.4 Future directions ........................................... 215

153
5.1 SUMMARY OF THE CHAPTER

In this chapter, I explore the Staufen family of double-stranded RNA binding proteins and their interactions with RNA structures, in particular duplexes that are bound by them in vivo. This chapter is divided into two parts. First, I consider Staufen1 (STAU1) in HEK293 cells. Here, I use the extended atlas of STAU1-bound duplexes obtained using data from Sugimoto et al. (2015) using my updated analysis method as detailed in Chapter 4. In order to gain insight into the nature of STAU1 binding, I aim to contextualise the duplexes bound within the universe of experimentally determined RNA duplexes. Then, I infer some potential functional effects of STAU1 binding, particularly in terms of RNA metabolism.

Second, I study Staufen2 (STAU2) in rat brain tissue. Here, I aim to develop the previous work in a more physiological context by assaying STAU2 bound structures in neuronal tissue using hiCLIP. I start by describing the general characteristics of the STAU2-bound duplexes we detect; in particular I highlight evidence supporting the validity of the computational hybrid detection pipeline. Then, I explore the changes in binding to structures across mammalian development genome-wide, before focusing on some specific neuronally-relevant targets. Finally, I examine potential interactions between STAU2 and other RBPs, using the literature to help guide the focus of our analyses.

Ultimately, with further integration with orthogonal functional data, we anticipate these studies will offer new insight both into the functions of RNA structures, notably those bound by RBPs, and also generate mechanistic insight into how RNA structures bound by Staufen contribute to neuronal processes though RNP assembly, RNA localisation and RNA stability. I conclude with suggesting future directions for this work.

The STAU2 work is a collaboration with Dr Flora Lee from the Ule Lab who performed all of the iCLIP and hiCLIP experiments presented here and Dr Sandra Fernández-Moya and Ms Janina Ehses from the Kiebler Lab who provided the rat brain tissue. I designed and performed all of the computational analyses.
As discussed in Chapter 1, interactions between RNA and associated trans-acting proteins are central to the post-transcriptional regulation of RNA molecules. Moreover, RNA does not simply form a linear structure, but rather complex conformations that are commonly adopted co-transcriptionally. These RNA structures can play an important role in interactions with proteins, most easily by regulating accessibility of the RNA to single stranded RBPs (ssRBPs). RNA secondary structure has been shown to be an important determinant of ssRBP binding in vivo (Stražar et al., 2016) as they cannot bind to paired RNA, but bind to particular sequence motifs (Dominguez et al., 2018). Conversely other RBPs that interact with RNA structures, such as Dicer, Staufen, protein kinase R, RNA deaminases, contain dsRBDs: regions of the protein that form an α-β-β-α structure that can interact with RNA largely without simple sequence specificity (Bycroft et al., 1995; Tian et al., 2004). Nevertheless, proteins such as STAU1 have been shown to exhibit specificity both in terms of the RNAs to which they bind, and where along the transcript they bind (Sugimoto et al., 2015). The nature of the specificity and the contributions of RNA structure and/or sequence in vivo remain unclear.

5.2.1 The Staufen family of RNA binding proteins

Staufen is a dsRBP, first identified in fruitflies (St Johnston et al., 1991, 1992) where it was found to be required for mRNA localisation in oocytes (Ferrandon et al., 1994; St Johnston et al., 1991). It is a highly conserved protein, with homologues to the fruitfly gene found in mammals, including mice, rats and humans (Buchner et al., 1999; Marión et al., 1999; Tang et al., 2001; Wickham et al., 1999), frogs (Yoon and Mowry, 2004), zebrafish (Bateman et al., 2004) and worms (LeGendre et al., 2013). In mammals there are two orthologues: STAU1 and STAU2. STAU1 is expressed in most cell types, while STAU2 is enriched particularly in the brain (Heraud-Farlow and Kiebler, 2014). The STAU1 gene has 10 different protein coding splice variants,
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

while STAU2 has 29 different splice variants, both protein coding and long non-coding RNAs. Generally, there are two common STAU1 protein isoforms and four STAU2 protein isoforms.

STAU1 and STAU2 both contain multiple dsRBDs (Figure 5.1). STAU1 contains 4 dsRBDs (having lost dsRBD1 that is present in Drosophila); dsRBD 3 and 4 are responsible for binding double-stranded RNA (Lazzaretti et al., 2018; Park and Maquat, 2013; Ramos et al., 2000; St Johnston et al., 1992). STAU2 contains all 5 dsRBDs that are present in *Drosophila*, although RBD5 is particularly degenerate. STAU1 and STAU2 are approximately 50% identical in terms of protein sequence (Buchner et al., 1999; Furic et al., 2008), with the functional dsRBDs 3 and 4 retaining greatest similarity: 78% and 81% respectively (Park and Maquat, 2013). It was thought that dsRBDs 1, 2 and 5 do not retain the canonical dsRBD RNA binding function of dsRBD 3 and 4, despite similar structures (Gleghorn and Maquat, 2014). This is true for dsRBD 5, but it has recently been shown that for STAU2, dsRBDs 1 and 2 display RNA binding activity of equal affinity and kinetics to dsRBDs 3 and 4, and that in fact the two pairs, 1–2 and 3–4 act in concert, binding to two stem loops (Heber et al., 2019). For STAU1, dsRBD2 and 5 do not appear to be involved in binding RNA, but possibly help fulfil Staufen’s functions through other protein-protein interactions: dsRBD5, for example, is involved in homodimerisation (Gleghorn et al., 2013). Mammalian Staufen proteins have also been shown to contain a tubulin binding domain (Wickham et al., 1999) and a Staufen-swapping

*Figure 5.1: STAU1 and STAU2 protein isoforms.*
Schematics of the protein isoforms of STAU1 and STAU2. Superscript numbers indicate the approximate mass (kDa) of the isoform; boxes represent functional and/or structural domains. RDB, dsRNA-binding domain; TBD, tubulin-binding domain; SSM, Staufen-swapping motif; aa, amino acids. Reprinted from Park and Maquat (2013), with permission from Wiley.
motif important for homo-dimerisation (Gleghorn et al., 2013).

The origin of the specificity of Staufen binding remains an unresolved challenge. Typically, there is no sequence specificity, but rather structural specificity, with early NMR studies showing a stem loop with 12 uninterrupted base pairs forming a preferred structure that optimally bound dsRBD3 of Drosophila Staufen (Ramos et al., 2000). Most structural insight has come from studies of Staufen dsRBDs in complex with known mRNA target binding sites, notably STAU1 and ARF1 (Figure 5.2) and STAU2 and Rgs4 (Heber et al., 2019; Kim et al., 2007; Lazzaretti et al., 2018). Subsequent analysis focused on the interaction between STAU1 and the ARF1 transcript deduced a critical 19 base-pair stem (Figure 5.2A) within a more complex structure in the intervening 100 nt loop of mRNA (Kim et al., 2007). The binding is driven predominantly by interactions with the sugar backbone, although some bases in the minor groove of the duplex structure also make contact with the protein, suggesting that there may be some sequence recognition (Lazzaretti et al., 2018; Masliah et al., 2013). However, similar structures have not been found in other Staufen targets. A
complementary, largely computational genome-wide approach using data predominantly from *Drosophila*, identified three types of Staufen recognised structures in 3′ UTRs, (Laver et al., 2013). Type 1 structures were unbroken stems of 10 to 15 in a contiguous segment, while types 2 and 3 allowed for bulges or short internal loops within the contiguous segment (Laver et al., 2013). Assessing for these structures in STAU2 targets from rat neuronal tissue found the type 3 structures (shorter with bulges) to be enriched (Heraud-Farlow et al., 2013).

Interestingly, recent structural work examining murine STAU2 bound to Rgs4 proposed a model of tandem binding to duplexes: dsRBD 2 binds dsRNA with moderate affinity and slides along the stem, bringing dsRBD 1 and the stem in apposition; when a suitable dsRNA structure is found, dsRBD 1 also binds resulting in a high affinity binding at that site (Heber et al., 2019). A similar model is suggested for dsRBD 3 and 4, with the effect that two suitably located sites are necessary for both pairs of dsRBDs in order for a stable STAU2-RNA complex to form, a possibly means of increasing specificity.

5.2.2 Staufen functions: localisation, stability, translation

STAU1 and STAU2 share a number of functions. I consider these first before discussing evidence for their divergent functions. Mammalian STAU1 has been shown to mediate a wide range of cellular functions, including embryonic stem cell differentiation (Gautrey et al., 2008), mRNA localisation (Bauer et al., 2019; Heraud-Farlow and Kiebler, 2014; St Johnston et al., 1991; Tang et al., 2001), nuclear export (Elbarbary et al., 2013), RNA stability (Gong and Maquat, 2011; Heraud-Farlow et al., 2013; Kim et al., 2005, 2007) and RNA translation (Dugré-Brisson et al., 2005; Micklem et al., 2000; Ricci et al., 2014; Sugimoto et al., 2015). Although Staufen has been implicated in a wide range of functions, here I focus primarily on three: RNA localisation (specifically in neurons), RNA stability and RNA translation.
5.2. INTRODUCTION

5.2.2.1 Localisation

There is a wealth of evidence for the importance of Staufen in RNA localisation starting with its early characterisation in Drosophila development as being necessary for the localisation of a number of critical mRNAs (Roegiers and Jan, 2000), including oskar (St Johnston et al., 1991) and bicoid (Ferrandon et al., 1994). These localisation roles are also important during neurogenesis (Heraud-Farlow et al., 2013; Li et al., 1997), for example, STAU2 forms a complex with the RBP Pumilio2 and mRNAs for b-actin and prox1 in mammalian neural stem cells, with perturbation of the complex resulting in premature differentiation and aberrant prox1 localisation (Vessey et al., 2012).

Furthermore, RNA localisation is also very important for the proper function of mature neurons. As discussed in Chapter 1, these are highly polarised cells with sub-compartmentalisation, that results in a proteome at the dendrite or axon that can be different from that in the cell body: a process that is driven by the need for local translation at the synapse (Holt et al., 2019). In order for localised translation to take place, mRNAs have to be transported from the nucleus in the cell body to the synapse. Staufen plays an important role in targeting selected mRNAs for trafficking (Tang et al., 2001) possibly through interactions with alternative 3′ UTRs that are known to be important for directing transcripts (Tushev et al., 2018). Retained introns may also be part of an important mechanism by which Staufen helps localise particular transcripts to dendrites in neurons. For both Camk2a (Ortiz et al., 2017) and Calm3 (Sharangdhar et al., 2017), STAU2 interacts with specific retained introns in a subset of transcripts of these genes to enhance dendritic localisation in an activity-dependent manner.

5.2.2.2 Stability

In some settings, Staufen has been shown to bind to particular 3′ UTR structures in target mRNAs leading to degradation via a process termed Staufen mediated decay (SMD) (Park and Maquat, 2013). These Staufen-bound structures are formed through pairing of Alu elements (the predominant member of the short interspersed

159
nuclear element, SINE, repeat family) usually between two different RNA molecules or species (i.e. intermolecular duplexes), for example IncRNA-mRNA (Gong and Maquat, 2011) or mRNA-mRNA (Gong et al., 2013). The SMD pathway feeds into the nonsense mediated decay (NMD) pathway with the helicase UPF1, which is recruited by either STAU1 or STAU2 dimers (Gleghorn et al., 2013; Kim et al., 2005; Park et al., 2013) and enhances its helicase activity (Park et al., 2013; Park and Maquat, 2013). Although the majority of earlier work suggested that UPF2 was not necessary for SMD, unlike the NMD pathway, (Gong and Maquat, 2011; Gong et al., 2013, 2018; Kim et al., 2005; Park et al., 2013; Park and Maquat, 2013), more recent structural evidence suggests that the Staufen dimer may be recruiting UPF2, and it is in fact UPF2 that recruits and activates UPF1 rather than Staufen (Gowravaram et al., 2019). Finally, although SMD has been found to contribute to myogenesis and adipogenesis (Park and Maquat, 2013), other studies have found different effects of Staufen on reporter mRNAs in different cell types (Miki et al., 2011) which, together with the lack of detection of intermolecular Alu duplexes when directly examining STAU1-bound duplexes (Sugimoto et al., 2015), raises the possibility that SMD may not be a universal mechanism.

5.2.2.3 Translation

There is both biochemical and genomic evidence that STAU1 directly associates with ribosomes (Luo et al., 2002; Sugimoto et al., 2015), localises to the rough endoplasmic reticulum (Wickham et al., 1999) and is involved in regulation of translation. Conversely, for STAU2 in neurons, mRNP complexes are associated with factors such as CBP80 and PABPN1, while eIF4E and the majority of ribosomal proteins are absent, suggesting a translationally repressive environment (Fritzsche et al., 2013).

However, the nature of this regulation is not fully characterised, with conflicting results from further studies directly examining STAU1 binding and ribosome profiling. mRNAs with STAU1-bound structures in the coding sequence (CDS) have lower levels of translation (as assessed by translation efficiencies) (Sugimoto et al., 2015). However, another study found that STAU1 associated with actively trans-
lating ribosomes in a manner that was proportional to GC content and propensity for secondary structure formation: in mRNAs with higher GC content in the CDS, higher STAU1 levels were associated with higher ribosome densities (Ricci et al., 2014), in other words higher levels of translation. One possible explanation for this discrepancy is that the two studies characterised STAU1 binding in different cellular compartments: total cell (Ricci et al., 2014) or cytoplasm (Sugimoto et al., 2015). Another is that there may be differences in isolating direct and indirect associations with mRNAs, due to the methods used to isolate STAU1-associated RNAs: tandem affinity purification, where the RBP-RNA interaction is not stabilised (Ricci et al., 2014) and UV crosslinking and immunoprecipitation, where it is stabilised (Sugimoto et al., 2015).

5.2.2.4 Functional differences between STAU1 and STAU2

So far I have considered STAU1 and STAU2 as broadly similar proteins functionally. In addition to possible differences in their roles in translation, however, one study used microarrays to examine the mRNAs associated with STAU1 and STAU2 RNP s in HEK293 cells. They found distinct but overlapping sets of targets: approximately 60% of the probe sets identified in the STAU1 mRNPs were not found in STAU2 mRNPs and conversely approximately 75% of the probe sets identified in the STAU2 mRNPs were not found in the STAU1 mRNPs (Furic et al., 2008). Furthermore, although both STAU1 and STAU2 are involved in mRNA degradation and mRNA transport (Heraud-Farlow et al., 2013; Park and Maquat, 2013), STAU2 can also stabilise a subset of its transcripts (Heraud-Farlow et al., 2013). There are other distinct roles: for example, STAU1, but not STAU2, is involved in forskolin-induced long-term potentiation, whereas STAU2, but not STAU1, is involved in metabotropic glutamate receptor mediated long-term depression (Lebeau et al., 2008; Park and Maquat, 2013). In summary, together with the differing tissue expression patterns, there are both common and divergent roles for the two proteins.
5.2.3 Experimental methods to assess RNA structure

Given the importance of studying RNA structure to understanding post-transcriptional regulation, over the last twenty years many methods have been developed to probe this challenging facet of RNA biology, both experimental and computational. Here, I focus on the more recent experimental approaches that endeavour to perform genome-wide measurements to derive general principles. There are broadly two complementary categories of high-throughput sequencing approaches to study RNA structures. I discuss each of these in turn and highlight their potential for alternative insights.

1. Methods that probe nucleotide accessibility (using either enzymes or chemicals)

2. Methods that capture RNA duplexes or structures (using proximity ligation)

5.2.3.1 Methods that probe nucleotide accessibility

Early approaches to studying nucleotide accessibility, such as parallel analysis of RNA structure (PARS) (Kertesz et al., 2010; Wan et al., 2013) used two different RNase enzymes: RNase V1 and S1 nucleases that cleave double- and single-stranded RNA respectively. By generating a ratio of V1 to S1 reads, a PARS score is calculated which informs on the double-strandedness of a nucleotide position. An advantage of this approach is that it can identify regions of alternative structures in different RNA molecules, when both enzymes are able to cleave at the same position (i.e. in separate molecules that then contribute to the overall PARS score). The main disadvantage, however, is that PARS assesses in vitro RNA structure: total RNA is isolated from cells, enriched for polyadenylated transcripts and then renatured in vitro before being subjected to one of the two nucleases (Wan et al., 2013). However, the cellular environment means that in vitro structures are often different from in vivo structures, particularly with co-transcription folding of RNA, and so to address biological questions robustly, in vivo methods are necessary.
Consequently, a number of *in vivo* structure probing methods have been developed that use chemicals to modify bases (Ding et al., 2014; Rouskin et al., 2014) or the sugar-phosphate backbone (Siegfried et al., 2014; Spitale et al., 2015). These methods leverage the fact the reverse transcriptase will either truncate at, or mutate, these modified nucleotide positions. So, similar to CLIP methods, identification of these events equates to the identification of single-strandedness as these positions.

Dimethyl sulphate (DMS) reacts with the Watson-Crick face of adenosine and cytosine and is the basis for the DMS-seq (Rouskin et al., 2014) and structure-seq (Ding et al., 2014). However, this cannot assess single-strandedness at uridines or guanosine, and so chemical interference of RNA structures (CIRS)-seq in addition uses N-cyclohexyl- N’-(2-morpholinoethyl)carbodimide metho-p-toluenesulphonate (CMCT) which does modify these bases (Incarnato et al., 2014).

SHAPE (selective 2′-hydroxyl acylation and primer extension) methods such as SHAPE-MaP (Siegfried et al., 2014) and icSHAPE (*in vivo* click selective SHAPE) (Spitale et al., 2015) use chemicals that interrogate all four bases by selective acylation of the 2′ hydroxyl of flexible or accessible bases: NAI and NAI–N₃, in particular have been found to have a high signal to noise ratio (Lee et al., 2017). Notably icSHAPE also enriches for SHAPE modified RNA fragments using biotin-streptavidin isolation to enrich the sequencing library for structured RNAs (Spitale et al., 2015). As DMS methods are limited by the nucleotide specificities, SHAPE methods are often considered superior (Lee et al., 2017; Weeks and Mauger, 2011).

There are three important considerations when using these *in vivo* structure probing approaches. First, in terms of analysis, the truncation-based approaches (icSHAPE) can only probe one position on a given molecule, whereas mutation profiling (SHAPE-MaP) can assess multiple positions on the same molecule, within the limits of RNA fragmentation and sequencing read lengths. As such, mutation profiling has the advantage of being able to dissect heterogenous structures within a population and potentially give insight into alternative structures or dynamics (Siegfried et al., 2014; Zubradt et al., 2017). The cost is that a greater depth of sequencing is required to identify mutations resulting from chemical modification confidently. Second, these methods can only identify single-strandedness, as
there are no chemicals to-date that can modify double-stranded positions. Double-strandedness is inferred, therefore from regions of absence of single-strandedness, but an important possible alternative for these regions is that they are sites of protein binding. Protein bound to RNA can protect a single-stranded base from modification (Bevilacqua et al., 2016). Thirdly, although regions of double-strandedness can be inferred, which two regions are binding partners and give rise to an RNA duplex cannot be identified. The complexity of structured RNA, with its long-range structures and alternative conformations cannot be fully studied with these approaches. To address this last consideration, methods that capture RNA duplexes are required.

5.2.3.2 *Methods that capture RNA duplexes or structures*

To capture RNA duplexes directly, and thus identify interacting segments within an RNA molecule, or interactions between two RNA molecules, all methods currently use proximity ligation at their core. Flanking single-stranded RNA is digested away, leaving the RNA duplex behind. The two interacting RNA fragment ends (or duplex arms) are then ligated together to generate a single hybrid molecule, such that it can be sequenced as one continuous hybrid read. A means of stabilising and enriching for the RNA duplexes is necessary so that *in vivo* structures are preserved through the stringent RNA isolation processes and form a good proportion of the sequencing library. There are two current approaches to this problem that provide complementary information:

1. Using UV crosslinking to identify RNA duplexes bound by an RBP

2. Using Psoralen crosslinking to identify RNA duplexes transcriptome-wide.

There are two dsRBP-bound proximity ligation methods that use UV crosslinking: CLASH (crosslinking, ligation and sequencing of hybrids) (Helwak et al., 2013; Kudla et al., 2011) and hiCLIP (Sugimoto et al., 2017, 2015), both derived from CLIP methods discussed in Chapter 3. In humans, CLASH has been used to study miRNA-mRNA interactions, by using AGO1 (Helwak et al. 2013), while hiCLIP has been used to study predominantly RNA-RNA interactions associated with STAU1 within ribosomal and messenger RNAs as discussed (Sugimoto et al., 2015).
CLASH was the first proximity ligation approach (Kudla et al., 2011). As for CLIP, UV-C irradiation is used to crosslink the dsRBP to the bound RNA duplexes. The RBP-RNA complexes are then immunopurified. After digestion of the unprotected single-stranded RNA ends, the two arms of the RNA duplex are proximity ligated. Finally, the hybrid RNA is reverse transcribed and sequenced as for CLIP methods (Helwak et al., 2013). For CLASH, the two arms of the duplex are deconvolved from the hybrid reads using the hyb bioinformatic pipeline (Travis et al., 2014). As discussed in Chapter 4 hiCLIP builds on insights gained from CLASH with the introduction of an additional linker adapter that is used as an intermediary in the proximity ligation. Thus in the sequenced hybrid read the two arms of the duplex are separated by an identifiable 19 ribonucleotide adapter sequence. The use of the linker ensures unambiguous delineation of the two arms in the hybrid read before alignment using the hiclipr bioinformatic pipeline (Sugimoto et al., 2017). A priori information (e.g. Ago1 binding miRNA-mRNA duplexes for CLASH (Helwak et al., 2013)) is therefore not necessary to identify dsRBP-bound RNA duplexes: an agnostic approach can be used.

One notable limitation of these methods is the requirement for high sequencing depth and complexity to assess the full range of structures. The efficiency of UV crosslinking is variable depending on the RBP, but is estimated to be approximately 1–5% (Darnell, 2010; Porter and Khavari, 2019). Moreover, due to inefficiencies in the hiCLIP method, only 1–2% of the sequenced library consist of the hybrid reads with a linker adapter (Sugimoto et al., 2015).

Moreover, the above methods are limited to generating genome-wide atlases of structures that are bound by an RBP; they do not identify structures that are not bound. To address this and identify RNA duplexes globally, three methods were developed contemporaneously that use the chemical psoralen and its derivatives to crosslink duplexes: PARIS (psoralen analysis of RNA interactions and structures) (Lu et al., 2018, 2016), LIGR-seq (ligation of interacting RNA and high-throughput sequencing) (Sharma et al., 2016) and SPLASH (sequencing of psoralen crosslinked, ligated, and selected hybrids) (Aw et al., 2016). In essence, they use the specific and reversible nucleotide crosslinker 4′-aminomethyltrioxsalen (AMT), with SPLASH...
utilising a biotinylated version. AMT intercalates in RNA helices, with a preference for staggered uridines (Cimino et al., 1985), and upon photoactivation with UV light at 365 nm crosslinks the two strands in vivo. The crosslinked RNA duplexes can be extracted and selected and then undergo proximity ligation without a linker adapter. Irradiation with UV light at 254 nm reverses the crosslinking, to generate a single-stranded hybrid molecule that can undergo reverse transcription and sequencing. The major difference between the three methods lies in their approach to crosslinked RNA duplex enrichment: in PARIS, this is done with 2D gel purification; in LIGR-seq, with RNase digestion and in SPLASH, with biotin-streptavidin isolation. Furthermore each approach has its own bioinformatic method to identify the duplex arms from the hybrid reads. Technically, all three methods are able to map both intramolecular and intermolecular RNA-RNA interactions globally, although LIGR-seq and SPLASH focus more on intermolecular ones.

Again the main limitation of these psoralen crosslinking approaches is sequencing depth, which is even more of an issue as the assay is not limited to structures associated with one dsRBP, but instead attempt to identify all structures, both bound by any dsRBP or unbound. All three have been applied in eukaryotic cells (mouse and human) and the depth is insufficient to map the full spectrum of structures (as discussed later), in particular to assess structural dynamics. Although alternate structures have been identified on selected highly abundant RNAs (Lu et al., 2016), global assessment has not been possible. This is compounded by the relatively low yield of crosslinked RNA duplexes in the sequenced library: only 2.5–6% in PARIS (Lu et al., 2016). Although biotin-streptavidin isolation could significantly enrich the sequencing library, biotinylated psoralen has limited cell permeability (hence cells are preincubated with a detergent in SPLASH (Aw et al., 2016). COMRADES (crosslinking of matched RNAs and deep sequencing) is a more recently developed method that improves yield to 8–9% using a cell-permeable azide modified psoralen derivative (psoralen-triethyleneglycol azide) (Ziv et al., 2018). Moreover, a matched control experiment where the crosslink is reversed before proximity ligation ensures specificity of the identified duplexes (Ziv et al., 2018). Consequently, alternative structures in the 10 kb Zika virus can be reliably identified with ultra-deep sequenc-
5.3. METHODS: STAU1

The processing of the data for STAU1 hiCLIP in HEK293 cells was processed using methods described in Chapter 4. Here, I detail only the methods specific to the further analysis done here.

5.3.1 Processing of PARIS data

Raw FASTQ files of PARIS data in HEK293 cells were obtained from SRA (accession numbers: SRR2814763, SRR2814765 and SRR2814765). Following the methods described in Lu et al. (2016), adapters were removed using Trimmomatic (Bolger et al., 2014) and PCR duplicates collapsed using the script readCollapse.pl from the icSHAPE pipeline (Spitale et al., 2015). Reads were aligned to the same mapping indices as for STAU1 hiCLIP using the four stage mapping as described in Chapter 4 for the non-hybrid reads (Section 4.3.2).

As there was no bimodal distribution of hybrid binding energies (but still a difference compared to shuffled controls), the Gaussian mixture model filter from Section 4.3.3 was not applied. Confident duplexes, i.e. duplexes with more than one hybrid supporting them, were selected using an updated version of the FindIslands function from the hiclipr package (Sugimoto et al., 2017) that enabled parallelisation to permit a reasonable runtime.

5.3.1.1 Binding energy analysis

The binding energy analysis was performed as in Chapter 4. In brief, RNAduplex from the ViennaRNA package (Lorenz et al., 2011) was used to anneal the two duplex arms with the command: RNAduplex -noLP < <left arm sequence>\n<right arm sequence> > <dot-bracket output> to calcu-
late the minimum free energy of hybridisation of the two duplex arms. The control shuffled sequences with preserved dinucleotide frequency were obtained using uShuffle (Jiang et al., 2008) and the command `uShuffle -seed 42 -k 2 -n 1 -s <duplex arm sequence>`. Custom R scripts were use as wrappers to run the external programs and extract the minimum free energy calculation for each duplex.

5.3.1.2 Conservation analysis

Conservation analysis was performed using custom R scripts. Duplexes were converted from transcriptomic to genomic coordinates. A fixed window was generated around the centre coordinates of each duplex arm. The hg19 PhyloP 100-way conservation track was obtained from UCSC as a BigWig file. The windows were then intersected with the corresponding regions to obtain the PhyloP score at each position, with the mean at each position taken to generate the metaprofiles.

5.3.1.3 Analysis of 4sU-seq data

RNA-seq and 4sU-seq data from Mukherjee et al. (2017) were obtained preprocessed with expression and metabolism rates from GEO (Accession: GSE84722) with some supplementary files kindly provided by Dr Neelanjan Mukherjee. The values for each gene were merged with the STAU1 duplexes using the Ensembl gene ID.

5.3.1.4 Analysis of polyA sites

The BED file of polyA sites mapped to the hg19 genome was obtained from the PolyASite database, version r1.0 (Gruber et al., 2016). The position of polyA sites was integrated with windows around the duplex arm centres in a similar manner as for conservation using custom R scripts.

5.4 RESULTS: STAU1

Using the 5,419 confident STAU1-bound duplexes in HEK293 cells from Chapter 4 we can start to integrate these data with other publicly available datasets to learn
what is particular to these STAU1-bound ones.

5.4.1 Contextualising STAU1-bound duplexes

PARIS identifies, or at least samples from, the global universe of *in vivo* RNA duplexes (Lu et al., 2016). Furthermore, three PARIS experiments were performed in HEK293 cells, permitting a more equitable comparison with STAU1 hiCLIP; these are the focus of the comparison here. This is particularly important as there were marked differences in the regional distribution of duplexes between HeLa and HEK293 cells (Lu et al., 2016). I re-processed the PARIS data using the same pipeline as the updated hiCLIP pipeline for non-linker hybrids detailed in Chapter 4. This unified analysis minimised the introduction of biases due to different bioinformatic methods, tools or annotations, in particular using the same approach as in hiCLIP to identify confident duplexes (Sugimoto et al., 2017, 2015). Overall, the sequencing depth of PARIS was much greater than for hiCLIP which a much higher number of hybrid reads, resulting in a greater number of confident duplexes recovered (although proportionally, the ratio of hybrids to confident duplexes was similar for both PARIS and hiCLIP).

I confirmed the validity and comparability of the duplexes identified using this approach for both hiCLIP and PARIS in two ways. First, I assessed how they related to computational folding predictions. As discussed earlier, computational algorithms are limited to structures that span approximately 100 nt (Lange et al., 2012). Nevertheless, this is useful as a guide as sufficient detected duplexes fall

1Examining the reproducibility of the PARIS duplexes identified here across three biological replicates, found that the majority of duplexes were unique to each replicate: of the 113,940 confident duplexes identified across all three replicates, only 1,198 (1%) were common to all of them (data not shown). This low proportion may partly be accounted for by defining the duplex as the longest stem within the hybrid upon computational annealing of the hybrid arms (Sugimoto et al., 2015). In reality, the structure likely contains stems, bulges and internal loops; the 5 nucleotide leeway allowed to consider two duplexes as overlapping may not be sufficient to compensate. However, more likely is the possibility that the deep sequencing used in PARIS is still not sufficient to capture the vast number and diversity of cellular RNA structures, particularly when highly abundant structured RNAs, such as rRNAs, will be overrepresented in the library. As a crude calculation, with 50,000 different RNAs and 100,000 duplexes, there are only 2 structures per RNA; this is likely far less than the truth.

2Lu et al. (2016) show that in HeLa cells the majority (approximately 50%) of detected structures are in 3′ UTRs, whereas in HEK293 cells the majority are in CDS (approximately 45%).
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAU1 PROTEINS

Figure 5.3: Validation and comparability of hiCLIP and PARIS duplexes
(A) The mean probability of a 3′ UTR position being unpaired, as calculated by LocalFold (Lange et al., 2012), is plotted as a metaprofile around the center of the duplex arms. Only duplexes in 3′ UTR are included; proximal and distal arms are considered together. For PARIS, duplexes in 3′ UTRs without STAU1-duplexes are excluded. Duplexes are grouped into four categories according to the hybrid span (i.e. linear distance between the two duplex arms).
(B) The mean PhyloP 100-way conservation score is plotted as a metaprofile around the center of the duplex arms. Only duplexes in the 3′ UTR are included; proximal and distal arms are considered together.
within this limit. To avoid confounding the results with regional difference in binding, given that STAU1 is primarily binds 3′ UTRs, I restricted my analysis to duplexes detected in this region, and for PARIS to transcripts that also had STAU1 binding. For the computational predictions, I took the sequence of the last 100 nucleotides of the CDS and the whole 3′ UTR and used this as input to LocalFold (Lange et al., 2012) to predict the probability of each nucleotide being unpaired. Figure 5.3A shows the metaprofile of pairing probability around the duplex arm centre grouped by hybrid span. First, this reaffirms the poor accuracy of computational prediction in sequences longer than about 100 nucleotides. Second, for both STAU1 hiCLIP and the three PARIS replicates, there is an area of low unpaired probability around the centre of the duplex arm (i.e. consistent with an RNA structure at the core of the duplex). This result is more pronounced for hiCLIP than for PARIS.

Second, I assessed sequence conservation around the centre of the duplex arms. I used the PhyloP score across 100 species as a measure of conservation at the single nucleotide level (Figure 5.3B). Again, I focused the analysis to 3′ UTRs bound by STAU1, which showed that there is a peak of increased sequence conservation compared to the surrounding background. This is more noticeable for STAU1 hiCLIP than for PARIS, but lends support to the premise that the detected duplexes are real and biologically meaningful.

Having established that the two different sets of data could be used to make meaningful comparisons, I considered the general characteristics of STAU1-bound duplexes. Both hiCLIP and PARIS hybrid reads are more stable than their controls generated by shuffling the hybrid sequences while preserving dinucleotide content (Figure 5.4A). Hybrid reads identified by STAU1 hiCLIP are more stable than those recovered by PARIS, with a 9.3 kcal/mol difference in median minimum free energy. Here, I have only compared hybrids that are located in the 3′ UTR to ensure a fairer comparison. Of note, the detected hybrid arm length was slightly longer for PARIS than for hiCLIP, with medians of 26 and 23 nucleotides respectively (Figure 5.4B).

Next, I considered the regions of mRNA transcripts in which the duplexes were located. As STAU1 hiCLIP only accesses the cytoplasmic compartment (Sugimoto
et al., 2015), I exclude intronic duplexes from the PARIS analysis. 79% of STAU1 hiCLIP duplexes are located in the 3′ UTR, compared with 26% for the global duplexes identified using PARIS (Figure 5.5A). The majority of PARIS duplexes were in fact located in the CDS (62%). This highlights a marked preference of STAU1 for binding duplexes in the 3′ UTR of mRNAs, beyond the regional distribution of RNA duplexes globally.

Finally, one of the key original findings was the long-range duplexes bound by STAU1 that could not be detected using computational prediction algorithms, but only discovered experimentally using hiCLIP (Sugimoto et al., 2015). I calculated the range of the structures for both datasets (Figure 5.5B). The bimodal distribution of hybrid spans (i.e. the linear sequence distance between the two duplex arms) is once again apparent, with one short-range peak and one long-range peak that is a bit less frequent. This is likely accentuated with the recovery of the protected short range structures as described in Chapter 4. However, the distribution for PARIS is unimodal, with the peak overlapping the short-range STAU1 hiCLIP peak. Although longer-range structures are detected by PARIS, the long-range peak is unique to STAU1 hiCLIP.
5.4. RESULTS: STAU1

![Graph](image)

**Figure 5.5: Regional distribution of duplex arms**

(A) The RNA species or regions in which the duplex arms are located for STAU1 hiCLIP and for PARIS. Reads mapping to the canonical rRNA sequence were pre-filtered. Intronic reads were excluded as the STAU1 hiCLIP experiment only accesses the cytoplasmic compartment.

(B) The distribution of spans of 3′ UTR duplexes, i.e. the linear distance in nt between the two duplex arms, termed loop length in Sugimoto et al. (2015)).

5.4.2 STAU1 binding and RNA metabolism

In order to gain correlative insight into the possible effects of STAU1 binding on RNA metabolism, I integrated the hiCLIP data with data from a study in HEK293 cells examining RNA metabolism (Mukherjee et al., 2017). They generated profiles of RNA production and maturation after metabolic labelling of RNA with 4sU for 5 different durations. Then they used INSPEcT (de Pretis et al., 2015) to infer synthesis, processing and degradation rates for each gene by comparing primary and mature RNA concentrations of 4sU-labelled RNA and total RNA.

The 5,419 confident STAU1-bound duplexes are found in 2,730 different genes. For 2,188 genes, they are exclusively in the 3′ UTR, for 400, they are exclusively in the CDS, and for 212 there are duplexes in both regions. In order to fairly apportion these last category, I calculated the ratio of 3′ UTR to CDS counts. Any with a two-fold over-representation of one region was assigned to that region. This gave a total of 414 CDS genes and 2,236 3′ UTR genes. To create a valid control comparison...
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

Figure 5.6: Expression levels of STAU1-bound transcripts
The distribution of gene expression is shown according to the presence of STAU1-bound duplexes in either the CDS or 3′ UTR. The gene copy number in HEK293 cells was calculated using spike-ins by Mukherjee et al. (2017).

group of genes, it was not sufficient to consider genes without any detected STAU1 duplexes. The bias in CLIP method towards detecting binding in more abundant transcripts (Chakrabarti et al., 2018) is apparent when comparing the expression levels of the bound with the unbound transcripts (Figure 5.6). I filtered out those unbound genes that have an expression level less than the 5th percentile of the expression of all the STAU1-bound genes to arrive at a more comparable control set of 7,286 genes.

Now examining the synthesis, processing and degradation rates (Figure 5.7) we can see that transcripts with structures bound by STAU1 have higher synthesis and processing rates, but lower degradation rates. This is more pronounced for transcripts with structures bound in the CDS than in the 3′ UTR. Given that the majority of STAU1 bound structures are in the 3′ UTR, which has known roles in regulating RNA degradation, I focus on this region. Transcripts with longer 3′ UTRs have higher degradation rates (Figure 5.8). Furthermore, if we plot the distribution of polyA sites (Gruber et al., 2016), around the duplex arms we can see that the distal arm is often located at the end of the 3′ UTR transcript (Figure 5.9). This is not apparent for the proximal arm, however, with a small dip upstream of the proximal arm centre where the corresponding distal arm would be located in the case of a
5.4. RESULTS: STAU1

**Figure 5.7: RNA metabolism rates of STAU1-bound transcripts**

RNA metabolism, as assessed using synthesis (Syn), processing (Proc) and degradation (Deg) rates is shown according to the presence of STAU1-bound duplexes in either the CDS or 3' UTR. The rates were calculated using INSPecT (de Pretis et al., 2015) on 4sU-seq time course data by Mukherjee et al. (2017).

**Figure 5.8: 3' UTR length and RNA degradation rates**

RNA degradation rates are shown according to 3' UTR length. Rates were calculated using INSPecT (de Pretis et al., 2015) on 4sU-seq time course data by Mukherjee et al. (2017).
Figure 5.9: PolyA sites around duplex arms

The distribution of experimentally derived polyA sites (from Gruber et al. (2016)) is plotted around the centres of the duplex arms. Only duplexes in the 3′ UTR are included; proximal in the left panel and distal arms in the right.

Together with the long hybrid spans, this suggests that STAU1 might be binding duplexes that span large stretches of the 3′ UTR, extruding the intervening sequence. To gain a measure of this I calculate an adapted circularisation score (Aw et al., 2016) for each 3′ UTR:

\[
\text{circularisation score} = \frac{\text{mean loop length}}{3′ \text{ UTR length}}
\]

The score can range from 0 to 1, with a score of 1 indicating that the whole 3′ UTR is spanned by the duplex. If we divide the circularisation score into quartiles, we can see that there is a negative correlation between degradation rate and circularisation quartile (Figure 5.10). Moreover, if we further divide the transcripts into quartiles based on 3′ UTR length, we can see that although higher circularisation scores in long 3′ UTRs are associated with a higher degradation rate than those in short 3′ UTRs, overall the pattern of a higher circularisation score correlating with a lower degradation rate is preserved across all four quartiles of 3′ UTR length.
5.5. DISCUSSION: STAU1

5.5.1 Differences between hiCLIP and PARIS are a combination of biological and technical effects

When comparing results from hiCLIP and PARIS experiments, although great effort has been made to standardise the bioinformatic method, the experimental methods have major differences that could contribute to the observed differences. It is therefore crucial to consider these carefully before assigning a biological interpretation of the results.

To start, the differences in the hybrid arm width likely arise due to technical differences in the two methods. Different RNase enzymes (RNase I versus S1 nuclease and ShortCut RNase III, respectively) and conditions are used to fragment the RNA. It is difficult to understand fully what underlies the different stabilities of the hybrids recovered using hiCLIP and PARIS. I consider three possible explanations for this observation. First, while it is possible different hybrid arm lengths in the two could contribute to different binding energies, we would expect the STAU1
hiCLIP with its shorter lengths to be penalised and result in higher minimum free energies. Second, psoralen crosslinking may stabilise weaker duplexes that would otherwise by disrupted by the cell lysis buffers in hiCLIP, where the single-stranded nucleotide adjacent to the the duplex anchors the RBP-RNA interaction through crosslinking, but the duplex itself is only stabilised by association with the RBP. So, in hiCLIP, less stable duplexes may be lost or under-represented. Third, STAU1 may bind more stable structures than those generally found in 3′ UTRs. The higher sequence conservation observed for STAU1 hiCLIP again, may reflect technical artefacts. With the large number of duplexes recovered by PARIS with a lower stability, some of these may be more transient structures and therefore less subject to selection pressure as opposed to the binding sites of a protein.

The difference in regional distribution, with a predilection for 3′ UTR binding, is likely to reflect the real preferences of STAU1. Functionally, there are many reports of specific cases of STAU1 function being driven by 3′ UTR binding (Ferrandon et al., 1994; Lee and Mayr, 2019; Park and Maquat, 2013). In addition although it is possible that PARIS has a bias for a short range structures, it is not apparent from the method from where this could arise. The long-range duplexes are probably preferentially bound by STAU1 out of the population of RNA duplexes of all ranges present in cells. Depletion of SNPs from long-range duplexes suggests that some disease-causing mutations could be explained by effects on these structures (Sugimoto et al., 2015). Moreover, they may have functional roles particularly in RNA degradation.

5.5.2  STAU1-bound structures may have roles in RNA metabolism

Although there is evidence for STAU1 promoting mRNA decay though SMD (Gong and Maquat, 2011; Kim et al., 2005, 2007), here we see that transcripts with STAU1-bound 3′ UTR duplexes have slightly lower degradation rates, compared with expression-matched unbound controls (Figure 5.7). General differences between the SMD studies, such as different cell lines and transcriptomic context of the binding, could possibly underlie these contrasting observations. The majority of earlier work
5.6. METHODS: STAU2

investigating SMD were not genome-wide (Kim et al., 2005, 2007) and so may reflect specific rather than general functions of STAU1. More recent genome-wide analysis associated this effect to binding with SINE elements (B or ID in rodents and Alu in humans) (Lucas et al., 2018), however, there was no enrichment of Alu binding in duplexes detected by STAU1 hiCLIP (Sugimoto et al., 2015). One speculative explanation is that only the cytoplasmic compartment has been accessed, whereas SMD may require Staufen to bind in the nucleus.

Although these findings are correlative, the relationship between circularisation score and 3′ UTR degradation (Figure 5.10) is intriguing, particularly in light of the location of polyA sites around the duplexes (Figure 5.9). It appears that the STAU1-bound structures partially negate the effect of 3′ UTR length on RNA degradation. However, it is not clear if this effect is due to the RNA structure or due to STAU1 binding. This could be addressed by performing the RNA metabolism experiments in STAU1 knockdown cells (although STAU1 may have multiple effects). If the effect was not abrogated, then disrupting the RNA structure using, for example, catalytically inactive dCas13 methods (Terns, 2018) in selected representative transcripts and assessing degradation would help to confirm the role of the structure.

5.6 METHODS: STAU2

5.6.1 The hiCLIP experimental method adapted to tissue

Developments to the hiCLIP experimental method described in Chapter 4 and a detailed analysis of the improved approach and its steps are discussed in Lee (2019). This revised version was used for the STAU2 hiCLIP experiments performed in rat brain (Figure 5.11). The important change for the purpose of data analysis was omission of the linker adapter and proceeding directly to proximity ligation after the ligation of the SeqRv adapter.
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

Figure 5.11: The hiCLIP in rat brain
The key steps of the updated hiCLIP experimental method are shown. The major change is the omission of the linker adapter and proceeding directly to proximity ligation after the SeqRv adapter ligation. Figure by Dr Flora Lee, reproduced with permission.
5.6. METHODS: STAU2

5.6.1.1 Extension of the Rattus norvegicus annotation

Rat rRNA subunit sequences were obtained from NCBI (NR_033176.2, NR_046238.1, NR_046237.1, NR_046246.1). For the transcriptomic sequences, the Ensembl 94 Rattus norvegicus annotation was obtained as a GTF file. Protein coding genes had the 10 kb downstream of the 3′ end of the annotated gene added as the extended 3′ UTR region. Any genes that extended beyond the ends of the chromosome were trimmed back. Genes annotated as rRNA were removed from the Ensembl annotated genes. The sequences were obtained for these regions and combined with the NCBI rRNA sequences to generate a custom transcriptome for alignment.

The regions of the sequences that corresponded to regions of the transcripts was annotated based on the longest transcript. For protein coding genes the longest protein coding transcript was selected (using the biotype as annotated by Ensembl). Ties were broken using the hierarchy of most exons > longest 3′ UTR > longest CDS > longest 5′ UTR. Any sequence regions that remained unannotated (i.e. were not part of the longest transcript) were annotated based on a hierarchy of 3′ UTR > CDS > 5′ UTR > intron > other.

5.6.2 hiCLIP data pre-processing and read alignment

The read alignment and hybrid identification method was inspired by previous work performed for CLASH (Travis et al., 2014). A custom python script was used to extract the split UMI from the 5′ end sequence and place it in the FASTQ header. Illumina sequencing adapters were trimmed with TrimGalore, a wrapper for Cutadapt (Martin, 2011).

A preliminary filtering alignment step was carried out to remove two types of read: i) reads that align across annotated splice junctions and were therefore spliced reads rather than hybrid reads and ii) reads that aligned as one continuous segment to the transcriptome and therefore were not hybrid reads. First the reads were aligned to the rn6 genome using STAR (Dobin et al., 2013) with an index that had been created using the Ensembl 94 annotation, using the options:
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUEN PROTEINS

Then a custom python script was used to parse the BAM file to remove any reads that had aligned to annotated splice junction using the jM tag. The BAM file was converted back to a FASTQ file using BEDTools (Quinlan, 2014). The remaining reads were then aligned to the genome again, using Bowtie2 (Langmead and Salzberg, 2012) with the options: -local -norc, and the resultant BAM file parsing using another custom python script to remove reads that mapped perfectly using the NM tag. The final filtered BAM file was converted back to a FASTQ using BEDTools (Quinlan, 2014).

The filtered reads were then mapped to the custom transcriptome using pblat (Wang and Kong, 2019), a parallelised BLAT algorithm based on the original BLAT (Kent, 2002). This was run with the options: -stepSize=5 -tileSize=11 -minScore=15 -out=blast8 (Travis et al., 2014). This identified multiple possible partial alignments for a given read.

5.6.3 Hybrid read identification and duplex delineation

The output BLAST8 format mapping file was then parsed with a custom R script. First reads with an e-value > 0.001 and more than 100 possible alignments were removed. Then for each read, only alignments that matched the lowest e-value for that read were kept. Reads where an alignment only left a 17 nt stretch unaligned in a continuous segment were filtered. A cartesian join was performed on the remaining alignments to identify all possible combinations of partial read matches. Combinations where there was a significant overlap in the two query regions (i.e. the two partial read regions being aligned) were removed. This was defined as 4 nt;
a small overlap was permitted as a few nucleotides at the end of each arm could be the same and therefore map to two locations (Travis et al., 2014). Similarly reads that had an overlap in the subject regions (i.e. the two transcriptomic regions mapped to) were removed if they mapped to the same gene. Reads that were too far from the crosslink position (5 nt) were removed. Finally this set of reads with possible valid hybrids were split in two. Those with only one remaining combination were kept as single hybrid reads, and the rest as multiple hybrid reads. For each multiple hybrid read, the combinations were assessed to see if they corresponded with a hybrid present in any of the unique hybrid reads and if so that was selected as the hybrid for that read. The remaining multiple hybrid reads were discarded. Confident duplexes were identified as before as those with at least two supporting hybrid reads using an updated parallelised version of the FindIslands function from hiclipr (Chakrabarti et al., 2018).

5.6.4 Gene ontology analysis

Gene ontology analysis was performed using the clusterProfiler R package (Yu et al., 2012), using the Ensembl IDs of genes with STAU2-bound duplexes. The universe of all genes was selected as background. A p-value cut off of 0.01 and a q-value cut off of 0.05 was used.

5.6.5 Processing of rat brain RNA-seq data

Rat brain RNA-seq data from the rat transcriptomic BodyMap (Yu et al., 2014) was obtained at four different timepoints (P2, P6, P21 and P104). Raw FASTQ files, four replicates for each time point, were downloaded from SRA, combining technical replicates (Accessions: P2, SRR1169959–66; P6, SRR1169967–74, P21; SRR1169975–92; and P104, SRR1169983–91). Illumina sequencing adapters were trimmed using

---

4 Initially, for the remaining multiple hybrid reads, a representative hybrid was selected based on the one with the largest proportion of the read aligned, and the one with the shortest gap in the query sequence between the two arms, but ultimately these hybrids had low binding energies and therefore considered as spurious.
Cutadapt (Martin, 2011). Reads were aligned using STAR (Dobin et al., 2013) to the index generated in Section 5.6.2 using the standard ENCODE RNA-seq STAR parameters.

5.6.5.1 Differential binding analysis

For the two sets of differential binding analysis: i) STAU2 iCLIP v. mock iCLIP and ii) STAU2-bound duplexes at different time points, the DESeq2 package was used (Love et al., 2014). Unless otherwise described, the default methods and options were used. An adjusted p-value of \(< 0.01\) was considered significant.

Count tables were generated from the BAM (RNA-seq) or BED (iCLIP and hiCLIP) files using custom scripts in R based on the GenomicAlignments package. For iCLIP the counts were over the extended annotation previously generated, but at the gene level. For hiCLIP the counts were over the duplex arms, but also including the flanking 25 nt either site; the counts for both arms were combined to give a count for the duplex as a whole. Crosslinks or reads that overlapped more than one feature were discarded, equivalent to HTSeq’s “union” mode (Anders et al., 2015).

For principal components analysis, reads were normalised using a regularised log transformation. Either the top 500 genes (for iCLIP) or the top 100 duplexes (for hiCLIP) were used.

For the iCLIP differential binding analysis, when estimating size factors the shorth function from the genefilter package was used as it gives better results for low count data. Genes that had fewer than 5 normalised counts in 3 or more samples (out of a total of 6) were filtered.

For the hiCLIP differential binding analysis, a likelihood ratio test was used with a full model of \(\text{assay} + \text{age} + \text{assay} : \text{age}\) and a reduced model of \(\text{assay} + \text{age}\), where \(\text{assay}\) corresponds to either the STAU2 hiCLIP, or the rat brain RNA-seq and \(\text{age}\) corresponds to the time points P0, P7 and 5 week. This therefore tests for the effect of the interaction term \(\text{assay} : \text{age}\), which equates to the ratio of ratios:
5.7. RESULTS: STAU2

\[
\text{assay : age} = \frac{\text{hiCLIP}_A}{\text{RNA-seq}_A} \times \frac{\text{RNA-seq}_B}{\text{hiCLIP}_B}
\]

Here \( A \) and \( B \) are two time points, e.g. 5 weeks and P0, so we can see that this tests for whether the fold-change between the two time points is different in hiCLIP compared with RNA-seq, thus controlling for the effect of RNA abundance.

5.6.6 Miscellaneous analyses

Here, I have detailed the analysis particular to the STAU2 hiCLIP data. Other methods, such as assessing binding energies and conservation scores are as described in Section 5.3. For the visualisation of minimum free energy structures, the RNAfold webserver was used (Gruber et al., 2008). For the conservation analysis, the relevant PhyloP-20way and PhastCons-20way tracks for rn6 were obtained from UCSC.

HuR and Rbfox1 iCLIP data were processed as detailed in Chapter 3 using the STAR index generated in Section 5.6.2. Crosslink counts were normalised to library size using a ‘per million’ calculation. No peak-calling was performed.

5.7 RESULTS: STAU2

The STAU2 study was performed in rats, to gain insight into its function in a more physiological setting. We used rat brain tissue at four post-partum developmental time-points (P0, P2, P7, and 5 weeks). To help contextualise these ages, and how they relate to humans, milestones of rat development are detailed in Table 5.1.
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

<table>
<thead>
<tr>
<th>HUMAN</th>
<th>RAT</th>
<th>DEVELOPMENTAL MILESTONES</th>
</tr>
</thead>
<tbody>
<tr>
<td>23–32 weeks gestation</td>
<td>P1–3</td>
<td>Oligodendrocyte maturation state changes Immune system development Establishment of the blood-brain barrier</td>
</tr>
<tr>
<td>36–40 weeks gestation</td>
<td>P7–10</td>
<td>Peak brain growth spurt Peak in gliogenesis Increasing axonal and dendritic density</td>
</tr>
<tr>
<td>12–18 years old</td>
<td>P35–49</td>
<td>Reduced synapse density, plateauing at adult levels Refinement of cognitive-dependent circuitry. Ongoing myelination; increasing white matter volume and fractional anisotrophy</td>
</tr>
</tbody>
</table>

*Table 5.1: Stages of post-partum rat development*

Equivalent human and rat development stages with particular milestones are described. From Semple et al. (2013), with permission of Elsevier.

5.7.1 Extension of the Rattus norvegicus annotation

As discussed in Chapter 1, alternative 3′ UTRs in the brain are particularly long, and this has important implications for neuronal function. Furthermore, as already described, Staufen has a preference for binding to sites in the 3′ UTR (Sugimoto et al., 2015). However, the rat reference annotations are generally poor, particularly when considering untranslated regions of genes. One example is given in Figure 5.12 of the Camk2a gene, an important neuronal transcript for synaptic plasticity (Miller et al., 2002). The Ensembl reference annotation only shows a short 3′ UTR. However, RNAseq coverage in rat brain at three different ages from the rat transcriptomic BodyMap (Yu et al., 2014) shows that transcription continues for nearly 3,000 nt after the annotated end. Moreover, this region of the transcript contains multiple STAU2 crosslinking events from our data. It is therefore crucial to augment the reference annotation to ensure potentially important 3′ UTR binding events are not missed. I considered two main options for addressing this issue: i) to map to the genome, and ii) to extend the 3′ UTR annotation.

With the hybrid reads, mapping to the genome poses two challenges. First, with the approach of using BLAT (Kent, 2002; Wang and Kong, 2019) to perform multiple partial matches for each read, aligning to the whole genome becomes
5.7. RESULTS: STAU2

Figure 5.12: Extension of the Rattus norvegicus 3′ UTR annotation
Genome browser view of the Camk2a 3′ UTR, showing the Ensembl reference annotation (blue). The top three tracks (black) show normalised RNA-seq coverage in rat brain at P2, P6 and P21 time points from Yu et al. (2014) annotation track. The bottom three tracks show raw crosslink coverage from biological replicates of STAU2 iCLIP (red). Both the RNA-seq and iCLIP signals in rat brain extend beyond the reference annotation.

computationally intractable, both in terms of run-time and also memory usage. Furthermore, this also massively increases the search space\(^5\), which also increases the chance of mapping errors and false positives when selecting partial matches as correct duplex arms. Therefore, we are restricted to aligning to the transcriptome and need to proceed with the second option.

Previous studies exploring binding in mouse brain that encountered similar limitations of the reference annotation extended the reference annotation by 10 kb to capture the unannotated regions (Weyn-Vanhentenryck et al., 2014). Although this distance is somewhat arbitrary, the main limitation to applying this approach here is the possibility of extending into the downstream gene, resulting in duplicate sequences in the alignment index, which can be obviated. Hence, we extended all protein coding genes by 10 kb to generate additional extended 3′ UTR region. This extended annotation has been used for the analysis presented here. I have developed a more refined data-driven approach using RNA-seq that is only partly implemented, but is discussed further later.
5.7.2 Characterising STAU2 binding in rat brain

First, I characterised general patterns of STAU2 binding, before focusing on binding to RNA duplexes. STAU2 iCLIP\(^6\) was performed using cortex at three post-partum developmental time points (P0, P7 and 5 week) in triplicate, with matched mock IP controls, also in triplicate (Table 5.2). In the mock IP conditions, UV irradiation is still used to crosslink RNAs to RBPs. However, the targeted immunoprecipitation step is omitted. Instead all RNA fragments that are crosslinked to an RBP are included in the sequencing library. Therefore, the mock IP experiments are a representation of the global binding pattern of RBPs to RNA more generally (with a bias towards

\(^{5}\)For comparison, exome sequencing corresponds to approximately 1.5% of whole genome sequencing, albeit generally not including non-coding RNAs.

\(^{6}\)These were in fact hiCLIP experiments, as they underwent the proximity ligation step. However, by using only the non-hybrid reads, the data obtained is equivalent to an iCLIP experiment as discussed in Chapter 4 and Sugimoto et al. (2015) and analysed as such.

### Table 5.2: Overview of STAU2 and mock IP iCLIP experiments in rat brain

<table>
<thead>
<tr>
<th>AGE</th>
<th>IP</th>
<th>REPLICATE</th>
<th>UNIQUE CROSSLINKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td>STAU2</td>
<td>1</td>
<td>557,295</td>
</tr>
<tr>
<td>P0</td>
<td>STAU2</td>
<td>2</td>
<td>268,350</td>
</tr>
<tr>
<td>P0</td>
<td>STAU2</td>
<td>3</td>
<td>321,424</td>
</tr>
<tr>
<td>P7</td>
<td>STAU2</td>
<td>1</td>
<td>362,435</td>
</tr>
<tr>
<td>P7</td>
<td>STAU2</td>
<td>2</td>
<td>408,881</td>
</tr>
<tr>
<td>P7</td>
<td>STAU2</td>
<td>3</td>
<td>371,550</td>
</tr>
<tr>
<td>5 week</td>
<td>STAU2</td>
<td>1</td>
<td>228,886</td>
</tr>
<tr>
<td>5 week</td>
<td>STAU2</td>
<td>2</td>
<td>481,631</td>
</tr>
<tr>
<td>5 week</td>
<td>STAU2</td>
<td>3</td>
<td>242,715</td>
</tr>
<tr>
<td>P0</td>
<td>Mock</td>
<td>1</td>
<td>374,723</td>
</tr>
<tr>
<td>P0</td>
<td>Mock</td>
<td>2</td>
<td>1,651,651</td>
</tr>
<tr>
<td>P0</td>
<td>Mock</td>
<td>3</td>
<td>276,764</td>
</tr>
<tr>
<td>P7</td>
<td>Mock</td>
<td>1</td>
<td>743,365</td>
</tr>
<tr>
<td>P7</td>
<td>Mock</td>
<td>2</td>
<td>368,260</td>
</tr>
<tr>
<td>P7</td>
<td>Mock</td>
<td>3</td>
<td>401,872</td>
</tr>
<tr>
<td>5 week</td>
<td>Mock</td>
<td>1</td>
<td>582,862</td>
</tr>
<tr>
<td>5 week</td>
<td>Mock</td>
<td>2</td>
<td>530,885</td>
</tr>
<tr>
<td>5 week</td>
<td>Mock</td>
<td>3</td>
<td>337,607</td>
</tr>
</tbody>
</table>

For comparison, exome sequencing corresponds to approximately 1.5% of whole genome sequencing, albeit generally not including non-coding RNAs.

These were in fact hiCLIP experiments, as they underwent the proximity ligation step. However, by using only the non-hybrid reads, the data obtained is equivalent to an iCLIP experiment as discussed in Chapter 4 and Sugimoto et al. (2015) and analysed as such.
5.7. RESULTS: STAU2

To get an overview of the binding patterns, I examined the regional distribution of crosslinks across annotated RNAs (Figure 5.13). Across all experiments, between ~40–50% of crosslink sites were located in regions annotated as intergenic, likely reflecting unnannotated RNAs in the Ensembl 94 reference annotation (data not shown). The stringent experimental purification steps make an alternative explanation of DNA contamination unlikely. There was a significant enrichment in 3′ UTR binding in the STAU2 IP conditions over the mock at all ages, reflecting the known preference for STAU2 to bind this region. Combining all the time points, 5.6% and 9.2% of annotated crosslinks were in the 3′ UTR or extended 3′ UTR respectively for the mock IP, compared with 21.8% and 18.1% for the STAU2 IP condition. This also re-affirms the importance of extending the reference annotation when analysing brain data sets. Interestingly, there also appears to be a relative decrease in STAU2 binding in the CDS region at the 5 week time point in comparison to the earlier P0 and P7 time points, from 10.5% and 15.7% at P0 and P7 respectively to 3.3% at 5 weeks. One could speculate that this might reflect a change in the functional roles of STAU2 binding: although 3′ UTR-associated roles such as stability and localisation are important at all ages, a role in translation may be less important at the later time point, when adult synaptic density has been achieved (Table 5.1).
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

Figure 5.14: Principal components analysis of iCLIP binding for STAU2 and mock IP
Principal components analysis (PCA) of the normalised count of crosslinks on mature mRNAs for STAU2 and mock IP conditions at the three different time points. The top 500 bound genes are used.

Next, I examined STAU2 binding on a transcript level. As we are particularly interested in post-transcriptional mRNA regulation in the neuron, and in particular the roles of STAU2 in localisation and stability, I focused on mature mRNAs by excluding intronic regions and ncRNAs. The greatest difference in our data set (PC1, accounting for 64% of the variance) is between the STAU2 and mock IP conditions, as expected, but there is also a minor difference between time points, even for the mock IP condition (Figure 5.14). This suggests that the global RBP binding profile changes with ageing. Consistent with the difference in regional binding (Figure 5.13), the 5 week time point is particularly different to the P0 and P7 time points. Hence, I examined each time point separately when assessing for significantly enriched transcripts.

I assessed STAU2 binding enrichment over the mock IP condition (Figure 5.15). By using the mock IP as the control comparison, we can assess which transcripts are particularly bound by STAU2 relative to general RNP complexes. Moreover, this will also control for differences in RNA abundance as the bias of iCLIP to detect binding in more abundant transcripts will be similar for both mock and STAU IP conditions. A number of important neuronal transcripts were consistently enriched in the STAU2-bound fraction, such as Camk2a, Rgs4, Rbfox1, Nnat and Mapt. These
Figure 5.15: STAU2 binding enrichment on mature mRNAs
Volcano plot showing the log2 fold-change in STAU2 binding to mature mRNA transcripts, comparing STAU2 IP (3 replicates) with mock IP (3 replicates) conditions at three time points. Significant transcripts were defined as those with an adjusted p-value of < 0.01 and an absolute log2 fold-change of 1 and are indicated in black. Selected important neuronal transcripts are labelled.
are transcripts that are known to be functionally regulated by STAU2 binding. Overall, using an adjusted p-value threshold of < 0.01 and a fold-enrichment of > 2 there were 564 STAU2-enriched transcripts at P0, 534 at P7, and 368 at 5 weeks. 37.7% of significantly enriched transcripts were also detected in a data set from older RNA immunoprecipitation microarray study (data not shown), also assessing STAU2-bound transcripts in rat brain, albeit at embryonic day 17 (Heraud-Farlow et al., 2013).

**5.7.3 hiCLIP identifies STAU2-bound duplexes in rat brain**

An extensive range of hiCLIP experiments were performed across four post-partum developmental time points (P0, P2, P7 and 5 weeks) and across different brain regions (cortex, hippocampus, cerebellum, striatum/thalamus, olfactory bulb and brain stem); using lysis buffers of mild and medium stringency, infra-red or barcoded L3 SeqRv adapters, and with STAU2 IP, an IgG control or a mock IP (Lee, 2019). Preliminary analysis showed that experiments performed with the mild lysis buffer and one set of barcoded adapters generated small library sizes, resulting in a very low yield of hybrid reads. These were not suitable for downstream analysis. I focused on four high-quality experiments in cortex across all four developmental time points, using the medium stringency lysis buffer to define an atlas of STAU2-bound RNA duplexes (Table 5.3).

Detected hybrids were largely reproducible, with the same hybrids recovered

---

Table 5.3: Overview of STAU2 hiCLIP experiments in rat brain

<table>
<thead>
<tr>
<th>AGE</th>
<th>LIBRARY SIZE</th>
<th>UNIQUE HYBRID READS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td>533,310</td>
<td>46,904</td>
</tr>
<tr>
<td>P2</td>
<td>1,358,958</td>
<td>109,442</td>
</tr>
<tr>
<td>P7</td>
<td>666,136</td>
<td>52,321</td>
</tr>
<tr>
<td>5 week</td>
<td>591,342</td>
<td>22,560</td>
</tr>
</tbody>
</table>

Library sizes and number of unique hybrid reads recovered for hiCLIP experiments performed using the medium stringency lysis buffer at four different ages of rat development. These experiments were used to detect STAU2-bound duplexes.
5.7. RESULTS: STAU2

![Graph showing STAU2-bound duplexes on the Camk2a 3' UTR.](image)

*Figure 5.16: STAU2-bound duplexes on the Camk2a 3' UTR.*

Genome browser view of individual hybrid reads at four time points on the Camk2a 3' UTR. Each thick box indicates a duplex arm, with the line inbetween linking the two arms. Below, in purple is the Ensembl annotated 3' UTR end, in black is the mRNA coverage region, in green the PhastCons 20-way conservation score.

In multiple, if not all time points, as exemplified by the Camk2a 3' UTR example shown in Figure 5.16. We recovered tens of thousands of hybrids with a median of 49,612 reads containing hybrids across the four experiments (Table 5.3).

I first examined the RNA biotypes that were bound by STAU2 (Figure 5.17A). A median of 36.5% of hybrid arms\(^7\) aligned to rRNA across the four time points. This is consistent with previous findings for STAU1 and its role in translational control (Ricci et al., 2014; Sugimoto et al., 2015). There was a particularly high proportion of rRNA at the P7 time point (62% of hybrid arms); however this was not evident for the iCLIP experiments presented in Section 5.7.2 (data not shown), and is likely

\(^7\)Each hybrid is counted twice — once for each arm.
Figure 5.17: Types of hybrids detected
(A) Relative proportions of RNA biotypes associated with the transcripts in which the hybrid arms were located across the four time points.
(B) Relative proportions of intragenic (blues, within the same RNA) and intergenic (reds, between two RNAs) hybrids at each of the four time points.

...to have arisen due to technical differences between the experiments. Interestingly there was an enrichment of small nuclear RNA (snRNA) binding at the early P0 and P7 time points (19 and 15% respectively) that was not apparent in the later time points. This suggests that there may be different roles for STAU2 in early and late post-partum development. Moreover, as snRNA is a nuclear RNA, there may also be different localisations of the STAU2 protein at the different time-points. Consistent with this is a similar enrichment of intronic binding at early time points compared to late (data not shown).

Overall, there was a predominance of intragenic hybrids, that is hybrids within the same RNAs\(^8\) (Figure 5.17B). This is largely driven by the increased rRNA binding, but there is a significant contribution from 3′ UTR binding, particularly at the later time points, supporting the hypothesis that STAU2 forms RNP complexes on neuronal 3′ UTRs (Figure 5.18A). The higher proportion of intergenic hybrids, that is hybrids between different RNAs, at the early P0 and P2 time points are likely due to the increased snRNA binding seen, with links between different snRNAs (Figure 5.18B). There is also a high proportion of intergenic hybrids between intron regions; one speculation is that this could reflect the binding of SINE elements\(^9\) as part of the

\(^8\)While it is possible that these hybrids are between different RNA molecules of the same transcript, here, as in Sugimoto et al. (2015), we consider it more likely that they form on the same RNA molecule.

\(^9\)Although rodents do not have the primate-specific \textit{Alu} elements usually associated with SMD, B1
5.7. RESULTS: STAU2

Figure 5.18: Regions linked by hybrids
(A) Bar plot show the regions linked by two hybrid arms at the four different time points. Only intragenic hybrids (within the same RNA) are shown, with linked regions with fewer than 500 hybrids excluded.
(B) As for (A) but for intergenic (between two different RNAs) hybrids.

As in Section 5.7.2, we next focused on mature mRNAs to assess cytoplasmic binding and the role of STAU2 in RNA localisation and stability. Here, the patterns of regional binding were similar across all the time points (Figure 5.19A). The vast majority of duplexes were located in the 3′ UTR, with an average of 54% in the annotated 3′ UTR and 35% in the extended 3′ UTR. This again emphasises the importance of augmenting the reference annotation. Although binding to the CDS was much less overall, there was a slight reduction in the proportion with age also, similar to changes seen in the iCLIP (Figure 5.13) and possibly reflecting less of a SINE elements have been seen to fulfil similar roles as Alus for other RBPs, such as DHX9 (Aktaş et al., 2017).
role for STAU2 in translation.

The duplexes observed in the 3' UTR are long-range structures spanning an average of $\sim 500$ nt in linear sequence space (with duplexes in the extended 3' UTR region spanning longer distances than those in the annotation: medians of 571 and 432 nt respectively) (Figure 5.19B). The longest observed 3' UTR duplex spanned nearly 10 kb. These results are consistent with previous observations for STAU1 in HEK293 cells (Sugimoto et al., 2015). Moreover, it suggests that the short-range stem loop in ARF1 bound by STAU1 and the basis of important structural work (Lazzaretti et al., 2018) is only part of the spectrum of duplexes bound by STAU2 as evidenced structural work with murine STAU2 and Rgs4 (Heber et al., 2019).

Finally, to evaluate the possibility of STAU2 compacting the 3' UTR, in a similar manner to the circularisation discussed in Section 5.4.2, I examined the sequence around the duplex arms for the presence of the canonical polyA signal “AAUAAA” (Proudfoot, 2011) (Figure 5.20). As we have extended the 3' UTR sequences by 10 kb, the ends no longer accurately define the 3' ends of the transcripts, hence I use the polyA signal as a proxy. Although only an approximation, this nevertheless shows an enrichment of the polyA signal around the distal duplex arm, compared to the proximal, supporting a model of compaction from the 3' end.
5.7. RESULTS: STAU2

Figure 5.20: The canonical polyA signal around STAU2-bound duplexes
The mean coverage of the canonical polyA signal, “AAUAAA” is shown around the confident 3’ UTR duplex arms. The mean coverage is smoothed using a rolling mean with a 25 nt window.

5.7.4 The validity of the detected STAU2-bound duplexes

The broad consistency with previous studies in terms of the regions bound by the duplexes, suggests that the identified intragenic mRNA duplexes are real, and not as a result of artefacts, particularly in the computational delineation of the proximity ligation ‘join’ in the sequencing read. However, there are three other pieces of evidence that also support this conclusion.

First, I assessed the stability of the duplex arms by computationally predicting the minimum free energy from their hybridisation using an algorithm that permits G-U wobble pairing in addition to canonical base pairing (Figure 5.21A). As a control comparison, I shuffed the sequences of each of the arms, whilst preserving dinucleotide content. As discussed in Chapter 4, maintaining the dinucleotide frequency is important when defining a background signal to assess thermodynamic stability of RNA secondary structures (Babak et al., 2007). Duplexes at all four time points are significantly more stable than their shuffled counterparts, with an overall mean of $-23$ kcal/mol compared with $-14$ kcal/mol. Furthermore this is very similar to that observed for duplexes detected in hybrids reads with a linker in Chapter 4, where the controlled hybrid ligation reaction ensures unambiguous delineation of the duplex arms.
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

Figure 5.21: The validity of detected duplexes
(A) The binding energy, calculated as the minimum free thermodynamic energy of computationally hybridising the sequences for the two duplex arms is shown for intragenic duplexes in mature mRNA for the four time points. These are compared with the binding energies for shuffled sequences where dinucleotide frequencies have been preserved.
(B) Gene ontology analysis of biological processes enriched in genes with STAU2 duplexes.

Second, I performed gene ontology analysis of the mRNAs in which STAU2-bound duplexes were reliably detected to identify significant biological processes (Figure 5.21B). Unsurprisingly, given that we are examining brain tissue, all of the significant terms are neuronal. However, the three most significant terms include synapse organisation, cell morphogenesis and dendrite development, all of which relate to known neuronal functions of STAU2 (Heraud-Farlow and Kiebler, 2014).

Third, I examined sequence conservation, as measured using the PhyloP 20-way score across 20 vertebrate species (Pollard et al., 2010), around the duplex arms (Figure 5.22). This score measures evolutionary conservation at individual positions across 20 vertebrate species. Here, for both proximal and distal duplex arms, mean conservation peaks at the centre of the arm. This reinforces the interpretation that we are detecting real RNA duplexes, and moreover this argues that these duplexes are functionally relevant.
5.7. RESULTS: STAU2

Figure 5.22: Conservation around STAU2-bound duplexes
Sequence conservation, as measured using the PhyloP score across 20 vertebrate species, is plotted as a metaprofile around the centre of duplex arms from intragenic duplexes in 3′ UTR regions. Proximal and distal arms are analysed separately, with data shown for each of four time points. The mean score is smoothed using a rolling mean with a 25 nt window.

5.7.5 The dynamics of STAU2-bound duplexes

In addition to the reproducibility in the detection of duplexes we can see on inspecting STAU2 binding in the genome browser, it is also notable firstly that some duplexes demonstrate more binding than others, and secondly that there appear to be systematic changes with age: an increase at each time point in case of the Camk2a transcript (Figure 5.16). Despite experimental improvements to the hiCLIP method, it remains a largely qualitative assessment of binding, predominantly due to the low yield of hybrid reads as a proportion of the sequencing library. With iCLIP however, we can perform more quantitative comparisons of differential binding. Therefore, I integrated the two data sets to generate quantitative insight into the behaviour of STAU2-bound duplexes across three developmental timepoints. (P2 was omitted as we did not have the iCLIP data, but from the observations so far, its behaviour would likely be similar to P0). By visualising both the hiCLIP-detected duplexes and the iCLIP-detected crosslinks together (Figure 5.23) we can clearly see that the duplexes link together regions of reproducible crosslink signal. Moreover there are cases where duplex binding increases with age, such as Camk2a, and cases where it decreases, such as Nnat.
Figure 5.23: Examples of STAU2 binding to RNA duplexes across development
Two examples of 3’ UTRs with differential RNA duplex usage across development: increasing in Camk2a and decreasing in Nnat. Circos plots are shown for each of the three time points. From the outside in, the tracks show: i) normalised RNA-seq coverage ii) three replicates of normalised STAU2 crosslink coverage and iii) PhastCons 20-way conservation. Links show confident duplexes, with darker shades of blue indicate more binding/usage.
I pooled all the hybrid reads across all four timepoints from the hiCLIP experiments and, from this complete set, identified 3,651 confident 3′ UTR unique duplexes that were bound by STAU2 with at least two hybrid reads supporting them. Then to quantify the duplexes, I counted how many of the crosslinks from the iCLIP experiments were located in the duplex regions. To check the validity of this approach, I perform principal components analysis of the duplex counts to compare with that using the STAU2 crosslinks at the transcript level earlier (Figures 5.14 and 5.24). As before, the three replicates at each time point cluster together, and the majority of the variability (PC1, 89%) is explained by the 5 week time point clustering away from the P0 and P7 time points. There is much less difference between the P7 and P0 time points (PC2, 5% of the variance explained).

One major possible confounding factor is gene expression. Given that CLIP methods are biased towards detected binding events in more abundant transcripts, the differences observed may simply reflect changes in gene expression. To account for this, I mapped the RNA-seq data from the rat transcriptomic BodyMap (Yu et al., 2014) to Rn6 with my extended genome annotation to quantify the expression of each duplex. I then used DESeq2 to test for differential structure usage. In particular, I used a likelihood ratio test to assess whether the iCLIP crosslink enrichment above the RNAseq expression was different at P7 compared to P0 and at 5 weeks compared to P0. Here, in the model I not only included terms for age and duplex quantification.
using iCLIP or RNA-seq counts, but also an interaction term for the two, which equates to this relative fold-change.

Even after accounting for differences in expression, 412 out of 3,651 confident 3’ UTR duplexes are differentially bound between the 5 week and P0 time points, using an adjusted p-value cut off of 0.01 (Figure 5.25). 263 show increased usage, while 149 show decreased usage. Many of the genes in which these duplexes are located are known to be important for neuronal development and plasticity, such as Camk2a.

If we look at how the normalised counts change across the different ages, and use hierarchical clustering to identify groups of structures that behave similarly, we can see that there are four main patterns of changing STAU2-binding to duplexes (Figure 5.26); the two largest groups are 1 and 3, and the smallest is 4.
5.7. RESULTS: STAU2

Figure 5.26: Heatmap of duplex counts at different developmental ages
(A) The log$_2$ fold change in STAU2-bound duplexes at P7 and 5 week over P0
(B) The normalised counts of STAU2-bound duplexes for each of the triplicates
at each of the three time points for differentially used RNA duplexes. Each
row is a STAU2-bound RNA duplex, and is centered and scaled. Hierarchical
clustering has been performed using the Ward D2 method.

1. Duplexes that are highly bound at P0 and decrease through P7 and 5 weeks.
2. Duplexes that are lowly bound at P0 and increase through P7 and 5 weeks.
3. Duplexes that are lowly bound at P0 and P7 and increase at 5 weeks.
4. Duplexes that are lowly bound at P0, increase binding at P7 and then
decrease at 5 weeks.

5.7.6 Structures of STAUI2-bound duplexes

Next, I examined the types of structures formed by the STAUI2-bound duplexes
for the highly bound duplexes in the Camk2a and Nnat 3’ UTRs (Figure 5.27). I
used RNAfold (Lorenz et al., 2011) to calculate the minimum free energy structures
formed by hybridising the two duplex arms. These structures were stable and
largely consisted of $\sim$12 nt continuous stretches paired nucleotides, with small
Figure 5.27: Structures of STAU2-bound duplexes
(A) Above, the RNAfold minimum free energy structures are shown for the three highly bound duplexes in the Camk2a 3' UTR. The colour scale indicates pairing probability. Below, the location of all the confident unique duplexes with the highly bound three indicated in red.
(B) As for (A) but for the two highly bound duplexes in the Nnat 3' UTR.
### 5.7. RESULTS: STAU2

<table>
<thead>
<tr>
<th>AGE</th>
<th>IP</th>
<th>UNIQUE CROSSLINKS</th>
<th>DUPLICATION RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td>HuR</td>
<td>8248432</td>
<td>1.62</td>
</tr>
<tr>
<td>P7</td>
<td>HuR</td>
<td>24787716</td>
<td>1.58</td>
</tr>
<tr>
<td>5 week</td>
<td>HuR</td>
<td>7721469</td>
<td>1.57</td>
</tr>
<tr>
<td>P0</td>
<td>Rbfox1</td>
<td>365406</td>
<td>2.03</td>
</tr>
<tr>
<td>P7</td>
<td>Rbfox1</td>
<td>566008</td>
<td>2.08</td>
</tr>
<tr>
<td>5 week</td>
<td>Rbfox1</td>
<td>225241</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Table 5.4: Overview of HuR and Rbfox1 iCLIP experiments in rat brain

Unique crosslink counts after collapsing PCR duplicates are shown for HuR and Rbfox1 iCLIP experiments in rat brain at the three different time points as STAU2 hiCLIP. The PCR duplication ratio is shown as one measure of data quality.

(< 3 nt) internal loops or bulges. In two-dimensions, the structures were mostly symmetrical. Interestingly, the two examples offer two different methods of 3′ UTR compaction. For Camk2a there could be a sequential looping of of intervening sequences in a concertina-like effect, while for Nnat the two highly-bound duplexes form a loop-within-a-loop conformation. This is also apparent in the Circos plots of the structures (Figure 5.23).

### 5.7.7 Interactions with other RBPs

One interesting observation from Figure 5.22 and also notable in the examples in Figure 5.23 is that the conserved region around the centre of the duplex arm appears to extend beyond the region of direct STAU2 binding by up to 20–30 nt either way. Not only does this suggest that these duplex binding sites are functionally important, but also we hypothesised that they may contain binding sites of other RBPs that interact with Staufen to form functional RNPs.

As a preliminary exploration of this hypothesis, iCLIP was performed for HuR (an ssRBP that has STAU1 associated functions in 3′ UTR dependent protein complex assembly (Lee and Mayr, 2019)) and Rbfox1 (an ssRBP that has important functions in splicing regulation (Lovci et al., 2013)), notably in brain development (Weyn-Vanhentenryck et al., 2014)). Only single experiments in rat brains at each of the three time points have been undertaken so far (Table 5.4). Sequencing libraries
for both RBPs were of good quality, as evidenced by the low duplication ratios. However, for Rbfox1, the sequencing depth, and so the number of unique crosslinks, was over an order of magnitude lower than HuR. This means that genome-wide patterns are likely to be more difficult to resolve for this RBP.

Examining the regional distribution of the crosslinks (Figure 5.28) shows a broad enrichment for 3' UTR binding, particular at the 5 week time point, where 60% of the HuR crosslinks and 71% of the Rbfox1 crosslinks were located in the 3' UTR.

Looking at individual examples, we can see there are potential relationships between STAU2-bound duplexes and HuR or Rbfox1 binding sites (Figure 5.29). In Camk2a there are two modes of interactions seen: i) binding either side of a duplex, bringing together two HuR and Rbfox1 binding sites that are over 500 nt apart, suggesting that multivalency of HuR and/or Rbfox1 binding is enabled by the STAU2-bound duplex; and ii) binding in the loop region extruded between two duplex arms.

I then explored these possible interactions genome-wide using RNA maps (Chakrabarti et al., 2018; Witten and Ule, 2011) to characterise the binding landscape around STAU2-bound duplexes. First, I calculated the mean sequence conservation (mean PhyloP 20-way score) of each confident duplex, and categorised them into

![Figure 5.28: Regional distribution of HuR and Rbfox1 binding](image)

Bar plot showing the distribution of all annotated crosslinks from iCLIP using HuR (left) and RBfox1 (right) IP. eUTR3, the extended 3' UTR region.
5.8. DISCUSSION: STAU2

5.8.1  hiCLIP detects valid RNA duplexes in brain tissue

We have identified thousands of duplexes bound by STAU2 in rat brain. Similar to STAU1, STAU2 has a strong preference for binding structures in the 3′ UTR...
Figure 5.30: HuR binding around STA2-bound duplexes
RNA maps showing HuR crosslinks around STA2-bound duplex arm centres for the three time points. HuR crosslink signal is normalised for library size. Duplexes are categorised by conservation, with the top quartile (Q4) of conserved duplexes considered high, and the bottom quartile (Q1) considered low.
5.8. DISCUSSION: STAU2

Figure 5.31: Rbfox1 binding around STAU2-bound duplexes
As for Figure 5.30 but for Rbfox1.

of transcripts. Furthermore, many of these structure are long-range, with mean
duplexes spans of many hundreds of nucleotides. Given the predominance of long 3’
UTRs in the brain (Lianoglou et al., 2013), and the limited rat reference annotations,
it was necessary to augment the annotation when analysing our data.

An alternative, more data-driven approach that it is partially implemented, is
to use RNA-seq data in these tissues to assemble transcripts de novo to supplement
the reference annotation. We can use a tool such as StringTie (Pertea et al., 2015) to
assemble the RNA-seq reads into reliable transcripts or fragments. Usually a high
sequencing depth is necessary for de novo transcriptome assembly, but as we can
select those fragments that are contiguous with and extend annotated 3′ UTRs, this may not be as strict a requirement. This will also accounting for signal drop-out due to difficulties in aligning to regions of low mappability. For this, it is important to generate our own tissue-specific data. By intersecting the annotated 3′ UTR with the overlapping extension we can generate an extended 3′ UTR region that encompasses fully the transcribed transcript with a data-defined 3′ end. This approach will be particularly important for future integration of our results with other orthogonal sequencing data.

The importance of this initial annotation extension step is highlighted both by the generally STAU2 binding patterns (Figure 5.13), but also by the large proportion of hybrid arms mapping to the extended 3′ UTR region (Figures 5.17). We have taken several angles to validate the duplexes we have detected by demonstrating: i) stable binding energies compared to shuffled controls; ii) preserved sequence conservation compared to the surrounding background; and iii) enrichment in gene ontology terms relevant to STAU2 biological processes. We also detected many reproducible rRNA-rRNA duplexes bound by STAU2. A further confirmation of the validity of our duplexes would be to overlay these with known rat rRNA structures as previously done for STAU1 hiCLIP and human 28S rRNA in Sugimoto et al. (2015). Finally, in accompanying mock IP conditions, minimal hybrid reads are recovered (data not shown).

It was interesting to find that there was a relatively higher proportion of snRNA duplexes and intronic duplexes at P0 compared with the later timepoints (Figures 5.17). One technical explanation for this is that physical disruption of brain tissue, and so the degree to which RNAs in the nuclear compartment are assayed, may be more difficult at later time points, which would suggest rather that there is an underestimation of snRNAs and intronic reads at P7 and 5 weeks. Generally, fewer hybrid reads are recovered at 5 weeks compared to the other time points (Table 5.3) which could reflect poorer RNA isolation. However, intronic reads are still detected at P7 and 5 weeks and the ratio of P7 or 5 week to P0 proportions is markedly different for snRNAs and introns (data not shown), which argues against variable accessing of the nuclear compartment. Another biological reason
could relate to STAU2 protein localisation at these different time points; its nuclear functions (Kiebler et al., 2005) may be more important at earlier time points, while functions on mature mRNAs, such as localisation more important at later ones.

For many transcripts, we detect multiple STAU2-bound duplexes. It is not possible to delineate from our data whether these structures occur on the same or different RNA molecules. Some of these structures are mutually exclusive, which suggests a degree of variability in STAU2 binding. Largely, however, we assume that non-redundant sets of STAU2-bound structures are present on individual molecules. This assumption leads to an extension of the 3′ UTR circularisation model discussed for STAU1 hiCLIP. We used the canonical polyA signal as a proxy for the polyA site and noted an enrichment immediately 5′ to the distal duplex arm. However, with the overall much richer dataset we have obtained using STAU2 hiCLIP, we hypothesise that these multiple structures lead to a compaction of the 3′ UTR, either through a concertina effect (e.g. Camk2a), or a loop-in-loop effect (e.g. Nnat). With the observation earlier that STAU1-bound long range 3′ UTR structures correlate with a lower degradation rate than would otherwise be expected for a transcript with a given 3′ UTR length, and previous data showing that transcripts in the brain have longer 3′ UTR isoforms (Lianoglou et al., 2013) that are necessary for the function of neuronal RNAs (Biever et al., 2019a; Glock et al., 2017), one could speculate the STAU2 has roles both in preventing degradation of the transcripts and in contributing to their functional roles, through localisation, for example.

5.8.2 STAU2 binding to RNA duplexes changes across development

There are marked differences in STAU2 binding to duplexes across development, exemplified in the cases of Camk2a and Nnat in Figure 5.23 and summarised in the PCA plot in Figure 5.24. Moreover, these changes are very reproducible (Figures 5.24 and 5.26). Given that RNA structures are commonly formed co-transcriptionally, it is not surprising that the RNA duplexes themselves remain the same across the three ages (Figure 5.16). There are four potential explanations for this observation of
CHAPTER 5. INSIGHTS INTO INTERACTIONS BETWEEN RNA STRUCTURES AND STAUFEN PROTEINS

dynamism.

1. STAU2 protein expression is different at the three time points.
2. RNA expression of the bound transcripts is different at the three time points.
3. STAU2 or RNA localisation is different at the three time points.
4. STAU2 binding to duplexes changes across development

From western blot experiments in neuronal cortical cultures, STAU2 has been found to increase in the first 5 days in vitro, plateau at high levels between days 5 and 10, and then slowly decline until the end of culturing at day 25 (Ms Janina Ehses, personal communication). Whilst difficult to link directly to rat brain, this would roughly correspond to an increases in STAU2 with neurogeneis, i.e. at embryonic time-points, reach a high plateau at post-natal days P0 to P7 and gently decline to a slightly lower steady state by 5 weeks. So, differences between P7 and P0 are unlikely to be due to changes in protein level. Furthermore, although this could contribute to a general reduction in STAU2 binding at 5 weeks compared with P0, the fact that we see opposing effects on binding on different transcripts, with some duplexes showing increased binding and others decreased binding, suggests that this is not a major contributing factor to the patterns we observe.

More likely is that changes in RNA expression as the rat ages results in changes in the numbers of duplexes that are available for STAU2 binding and hence also those that are detected by hiCLIP. Generally, the direction of change of binding to RNA duplexes on a given transcript is the same for all the duplexes would be consistent with these changes being primarily driven by changes in expression. In order to control for this confounding factor, I used RNA-seq data to quantify RNA expression for the STAU2-bound regions. This showed that even after accounting for gene expression, binding to the RNA duplexes changes over time (Figure 5.25). The publicly available rat BodyMap RNA-seq data (Yu et al., 2014) may not be representative of our rat brain data, and differences in precise rat genotypes, sampled brain regions, sampling time points, sample preservation methods and RNA extraction methods could contribute to inadequately controlling for true RNA expression in our rat brain samples. Furthermore, there may be differences in RNA
spliced isoforms, or polyadenylation site usage which could also contribute to these changes, but that remain unassessed with our data and approach.

Beyond the biological considerations, it is possible that the computational analysis method has limitations for our data types. Specifically, I have estimated a single dispersion estimate per duplex on the basis of both the iCLIP and the RNA-seq counts using DESeq2. The assumption is that the dispersion is similar for both data types, which may not be valid. Instead, it may be better to model the dispersions for the two data types separately, possibly by adapting methods developed for RNA methylation analysis (Liu et al., 2016, 2017) or by using an ensemble of approaches to define a confident set of dynamic sites (McIntyre et al., 2019).

Finally, it is worth briefly considering what further insights could be gained by comparing not only with RNA-seq, but also with the mock iCLIP. As this samples from the pool of RBP-bound RNAs, with a deeply enough sequenced mock iCLIP, this could control not only for RNA expression, but also for mechanistically relevant RNA expression. Structures that were dynamic after controlling for overall cellular RBP binding could indicate that the composition of RNPs associated with STAU2-bound structures were also changing.

5.8.3 Interactions with HuR and Rbfox1

The 3′ UTR of transcripts can act as scaffolds for RBP assembly: the CD47 gene has two alternative 3′ UTRs, with the long 3′ UTR bringing together SET and HuR RBPs to form a complex that is important for localising the translated CD47 protein to the plasma membrane rather than retaining it at the endoplasmic reticulum (Berkowitz et al., 2016). Recently, a study exploring the long BIRC3 3′ UTR found that cooperative binding of Staufen and HuR mediated its 3′ UTR-dependent complex formation that was necessary for its function in B cell migration (Lee and Mayr, 2019).

The extended regions of sequence conservation around STAU2-bound duplex arms could reflect selection pressures for binding sites for other RBPs that interact
with STAU2 to form RNP complexes. Alternatively, they could require those RNA duplexes or structures to form to enable their binding or for their functions. We selected two RBPs, HuR and Rbfox1, as potential interactors for a preliminary exploration of this hypothesis. Binding in the individual case study of Camk2a (Figure 5.29) supports this possibility, and so I investigated this further genome-wide. As our hypothesis originates from the observation of sequence conservation, I used the corollary that an absence of sequence conservation would correspond to an absence of RBP binding as the comparison for this analysis (Figures 5.30 and 5.31). Lowly conserved duplexes do not show much binding of either HuR or Rbfox, whereas there is generally higher binding around the duplexes for both RBPs for the highly conserved duplexes. All of these duplexes are confident, in that multiple hybrid reads support them, and follow the stable binding energy profiles in Figure 5.21A, so it is unlikely that false positive duplexes and contributing to this difference. A true negative control would be to perform iCLIP for a 3′ UTR binding RBP that we know does not associate with STAU2, but such a candidate has yet to be defined. Furthermore, around the highly conserved duplexes, there is a distinct profile to the RNA map with a peak of binding adjacent to the duplex arm centre. Intriguingly this is largely 5′ for both the proximal and the distal arms. This may reflect a bias in the iCLIP experimental method.

To establish the importance of STAU2 binding to these RNA duplexes for HuR or Rbfox binding, we plan to repeat the current experiments with replicates, and also in rat brains where STAU2 has been knocked down. If binding changes upon STAU2 knockdown, then this would support a model of STAU2 3′ UTR compaction enabling binding of other ssRBPs. Other methods in cell culture models, such as PARIS (Lu et al., 2016) or COMRADES (Ziv et al., 2018), would then be needed to see if the RNA duplex has also been affected upon STAU2 knockdown to ascertain if the structures are formed and stabilised, or merely bound, by STAU2.
5.8.4 Future directions

RNA-seq data generated in our rat brain samples will be invaluable to contextualise our findings and replace the rat brain RNA-seq data from BodyMap (Yu et al., 2014) that I have used here. Not only will this be useful for the most accurate extension of the rat brain transcriptome, but also it will enable us to control for RNA abundance optimally to identify STAU2 bound RNA duplexes that are dynamic across development, independent of changes in transcription, confirming our current results. Furthermore, it will enable us to perform analysis of alternative splicing and also polyadenylation. By obtaining these data at all three time points, we would be able to integrate changes in STAU2-bound duplexes to functional consequences. We know that for the Calm3 transcript, intron retention mediates the activity-dependent localisation function of STAU2, with peaks of STAU2 binding in this retained intron in the isoform that localises to dendrites (Sharangdhar et al., 2017). Genome-wide assessment of intron retention would help answer whether or not this is a general mechanism.

In the first part of this chapter, I discussed the concept of 3′ UTR compaction, which I assessed for STAU1 in HEK293 cells with the circularisation score. This is more difficult to apply to the STAU2 rat brain data set. Although a recent update to the PolyA_DB database (a comprehensive database of polyA sites identified using the 3′ end sequencing method 3′ READS) now includes rat (Wang et al., 2018a,b) this is a small dataset, and ideally characterisation of polyA sites in our rat brain samples using RNA-seq or a 3′ sequencing approach such as Quantseq would enable a robust assessment of the degree of STAU2-mediated 3′ UTR compaction through RNA structures.

Further the possibility of STAU2 binding SINE elements to direct Staufen-mediated decay remains to be explored in our dataset. At the earlier time points there are many intron-intron duplexes spanning different RNAs that could possibly be SINE elements (given their known enrichment in introns) and reflect activation of this evolutionarily conserved RNA degradation pathway (Lucas et al., 2018; Park...
and Maquat, 2013). A computational exploration for the presence of B SINE elements in the hybrid reads, or even an enrichment in the iCLIP binding sites would be useful.

To relate STAU2 binding to its role in neuronal localisation, I plan to integrate our data with orthogonal publicly available data for RNA localisation in rat neuronal compartments (Tushev et al., 2018). Although we have focused on brain tissue to gain more physiological insight into the functions of STAU2, in order to assess this further, we need to return to cell-based setups to enable manipulation of the system to help understand the mechanisms of STAU2 action. To assist in studying the localisation and stability roles we plan to fractionate cultured neurons into soma and neurites and use SLAM-seq (Herzog et al., 2017) to assess expression, 3′ end utilisation and stability. To complement this with a more reductionist approach, we plan to modify or delete selected STAU2-bound duplexes, from targets identified from the work presented, for example the duplexes in Calm3 and Camk2a, and assess the impact on STAU2 function. With these further experiments we aim to characterise the key regulatory roles of STAU2 in the RNA life-cycle in neurons.
CONCLUSIONS

6.1 Target-specific precision of CRISPR-Cas9 gene editing ............... 218
6.2 'Omes .................................................. 219
6.3 Single molecule genomics ........................................ 221
6.4 Integrative computational analyses ............................. 221
6.5 Final thoughts ............................................. 223
Throughout this thesis, I have presented many facets of computational approaches to investigate interactions between proteins and nucleic acids. In the first part, I focused on DNA; in the second, RNA. In Chapters 2 and 5, I presented analysis of data generated in close collaboration with experimental colleagues, showcasing the effectiveness of this crosstalk. In Chapters 3 and 4 I make use of different publicly available data sets and computational predictions to address methodological questions, demonstrating the possibilities to glean further understanding from already existing experimental results.

6.1 TARGET-SPECIFIC PRECISION OF CRISPR-CAS9 GENE EDITING

Despite a combination of DNA repair pathways, chromatin context and other variables, the editing outcomes at sites can be precise: with the same editing event recurring at high frequency. In the first part, I described our work investigating editing precision. We targeted over one thousand endogenous sites to characterise indel profiles. In so doing, we uncovered some remarkably simple rules that can be used to determine the kind of editing outcome to be expected at a site. In particular, the nucleotide at position $-4$ is key to precision and also whether or not there will be an insertion or deletion. By exploring the sequences around the DSB sites, we found that homology plays an important role in precision. In particular the identification of insertion homology led us to suggest that imprecision in the cutting by Cas9, through the generation of ragged ends, is the source of precision in editing outcome.

Two other studies that were published contemporaneously with ours (Allen et al., 2018; Shen et al., 2018) also converged on similar conclusions, highlighting both the robustness and also the relevance of our work. Beyond the research environment, as we build towards clinical applications of CRISPR-Cas9 it is imperative that we understand and can reliably predict the outcomes of editing and our work is a small step towards this aim. Machine learning methods and the use of much larger data sets will help in this goal. Our predictive model was trained on only hundreds of sequences, and was generated mainly as a tool to help understand the sequence
factors underlying precision. The overall predictive ability was modest. However, other recent predictive tools have been designed towards this primary goal with much larger training data sets and achieving better accuracies (Allen et al., 2018; Chen et al., 2019; Kim et al., 2019b; Leenay et al., 2019; Shen et al., 2018). Future work in this vein will only improve our capability to predict editing outcome.

Furthermore, the ability to edit precisely raises the possibility of targeting point mutations, or single nucleotide variants associated with disease directly, rather than the less efficient and more technically demanding option of having to replace a section of DNA containing the position of interest using pairs of sgRNA, engineered template sequence and homology-directed repair. It has recently been appreciated that using pairs of sgRNAs can induce complex rearrangements, usually unwanted (Kosicki et al., 2018). It is likely these were underestimated in the past owing to limitations in our ability to detect these, but the advent of long-read sequencing will make this a simpler prospect as it is more widely applied. Proof-of-concept experiments showing that CRISPR-Cas9 can correct gene defects have been undertaken (Park et al., 2016b; Shen et al., 2018) and indeed, in terms of somatic editing, a clinical trial using CRISPR-Cas9 to treat sickle cell disease has started recruiting patients. In parallel to advances in this technology, equally important are informed and wide-ranging discussions concerning the ethical implications, particularly in the context of germline editing.

6.2 ‘OMES

The genome, the transcriptome, the epigenome, the proteome, the methylome, the RBPome, the interactome…where next? The term genome was coined in the 1930s from combining ‘gene’ with ‘chromosome’ to mean “the sum total of genes in a set”. Nowadays, we are identifying, categorising and studying an ever wider range of ‘sets’. Even more, the number that is ‘the sum total’ is increasing with our improving technical abilities and ever advancing methods and technologies. In the RNA field, we are not satisfied with studying the sum total of RNA in
a cell, or even the sets in the nuclear, chromatin, or cytoplasmic compartments, which has been achieved for some years, but now are developing methods to study subcellular localisation of transcripts: for example sub-dividing the nucleus into nucleus, nucleolus, nuclear lamina and nuclear pore to study export from the nucleus (Fazal et al., 2019). Our understanding is growing of the role of membraneless cellular organelles and compartments, such as condensates formed through the aggregation of proteins and RNA physiologically, or under conditions of cellular stress, or through disease-causing mutations that affect the intrinsically disordered domains of RBPs.

The last ‘ome, the interactome, is a term that can encompass a great many interactions: it is not enough to study each ‘ome in isolation, but how they interrelate and interconnect is key to understanding biological systems. How the transcriptome interacts with the RBPome, and how can we study it has been the focus of the second part of my studies. I have taken a transcriptome-wide approach to study one particular family of RBPs, the Staufen family. While this has the possibility of generating some mechanistic and some functional insights, and is an important first step, understanding this two-way interaction will not be sufficient. Rather we will need to explore multi-way interactions, namely networks, as that is the true nature of biological systems. The reductionist approach enables us to study the constituent parts, and gain deep understanding about them, but the parts still need to be reassembled into the whole at some point. There is cooperativity in the actions of RBPs, and in this vein, I concluded Chapter 5 by analysing some preliminary data trying to integrate the STAU2 interactome with the HuR interactome and the Rbfox1 interactome. Moreover, cooperativity in the function of RNP complexes likely means that there is redundancy in the RBPs comprising them, and untangling essential from ancilliary functions within a network will be a challenge.
6.3 SINGLE MOLECULE GENOMICS

As alluded to earlier, sequencing methods are evolving with a new generation of long read sequencing methods. In one such method, Nanopore sequencing, the entire DNA or RNA molecule is fed through a pore, and the bases called by interpreting the electric current across the pore. This has meant that DNA over one megabase in length, and RNA hundreds of thousands of kilobases in length, has been sequenced in one go. The current challenge is a limited depth of sequencing.

In terms of RNA, this will revolutionise the splicing field, as transcript isoforms can simply be identified from sequencing, rather than complex computational stitching together of short spliced reads. Moreover, it may also be possible to leverage this technology to study RNA structure. One of our hypotheses is that Staufen is condensing the 3' UTR of transcripts, and that these conformations can bring together disparate binding sites. This facilitates repeated weak interactions, generating multivalent sites that may be necessary for the formation of functional RNP complexes. The hypothesis hinges on the assumption that the structures we have identified are on the same molecule; this needs to be challenged. Nanopore sequencing can be used to detect RNA modifications, such as m6A (Liu et al., 2019). If the bases were artificially modified to mark single-stranded positions, for example by using SHAPE chemicals, and if these could be detected by Nanopore sequencing, then we would be able to test our hypothesis by sequencing the entire the RNA molecule. Then, we could integrate these data with our hiCLIP atlas to relate these back to the STAU2-bound structures.

6.4 INTEGRATIVE COMPUTATIONAL ANALYSES

The publication of the first draft of the human genome was rightly hailed as a landmark in the nascent genomic revolution (Lander et al., 2001; Venter et al., 2001). At the time, Lander et al. (2001) concluded that, “In practice, our ability to transform such information into understanding remains woefully inadequate.”
Since that statement there have been major advances in computational biology and bioinformatic methods. While no longer, “woefully inadequate,” both the complexity of biological systems and the complexity of the data produced, means that deriving understanding remains an ever-present challenge.

Here, I have demonstrated the importance of integrative analysis. Different techniques provide insights into different facets of protein function. For example, in Chapter 3 by combining RNA-seq data, from which we can determine alternative splicing effects of an RBP, with CLIP data, from which we can identify binding sites of an RBP, we can derive mechanistic insight into the positional principles relating binding patterns to splicing outcomes. One major question in RNA biology at present is the formation and function of liquid-liquid phase separations or membraneless organelles; as touched upon in Chapter 1, these are likely important functions of proteins such as TDP-43 and may be the menas by which TDP-43 mutations cause ALS. Studying these as both the genomic level using technologies such as iCLIP and integrating these results both with functional genomic data about splicing or 3′ end processing will contribute insight to this area. Moreover, interpreting these finding with in light of those from other techniques, notably imaging will provide complementary data. Indeed, in the future, we will need to move beyond only integrating different genomic data, to considering the wider range of experimental modalities that shed light on biological systems. Here, for protein-RNA interactions, I have taken an RNA-centric view. There is also much work from the protein-centric view; and the two will converge resulting in a need to integrate genomic or transcriptomic data with proteomic data. One further exciting prospect is the combination of genomic with spatial or imaging data. New computational methods will be needed for this inter-disciplinary integration.

One of the aims of genetics is to understand the link between genes and disease. The omnigenic model builds on Fisher’s ‘infinitesimal model’ and proposes that, “gene regulatory networks are sufficiently interconnected such that all genes expressed in disease are liable to affect the functions of core disease-related genes,” (Boyle et al., 2017). Understanding these network relationships will be key to understanding complex traits and disease mechanisms and integrative analysis will be
essential to fulfil this goal.

6.5 FINAL THOUGHTS

Throughout this thesis, I have presented integrative computational analyses using both data from experiments designed in collaboration to address specific hypotheses, and publicly available data to complement or augment our interpretations and insights. The simple rules to predict CRISPR-Cas9 editing precision will, I hope, be useful to direct the use of this technology. The methods to assess CLIP data quality will be important as we start to assimilate databases of binding profiles for many RBPs. Fresh understanding of the original hiCLIP method driven by computational analysis has already led to experimental improvements, and helped enable its use in neuronal tissue. This has led to new insights into both spatial and temporal patterns of STAU2 binding and fresh perspectives into the role of this protein in the formation of RNP complexes. Interactions between proteins and nucleic acids; interactions between orthogonal data sets; and interactions between experimental and computational biologists will all be central to unravelling the complexity of biological regulatory networks that control the transfer and interpretation of the genetic code.


BIBLIOGRAPHY


BIBLIOGRAPHY


Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A.,


Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S.,


Queiroz, R. M. L., Smith, T., Villanueva, E., Marti-Solano, M., Monti, M., Pizzinga, M., Mirea, D.-M., Ramakrishna, M., Harvey, R. F., Dezi, V., Thomas, G. H., Willis,


BIBLIOGRAPHY


Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T.,


van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C.,


