Computational analysis of RNA-binding protein target-site selection and function

by

Xiao Li

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Molecular Genetics
University of Toronto

© Copyright by Xiao Li 2013
Computational analysis of RNA-binding protein target-site selection and function

Xiao Li

Doctor of Philosophy

Department of Molecular Genetics
University of Toronto

2013

Abstract

Gene expression is extensively regulated by the binding of RNA-binding proteins (RBPs) to cis-regulatory elements encoded in mRNA. A robust literature has emerged regarding the stability and intracellular distribution of transcripts genome-wide. However, clear binding preferences have only been reported for a few RBPs, and these binding data collectively explain only a small portion of these post-transcriptional events. I developed the RNA Regulatory Element Analysis and Discovery (RNA-READ) pipeline, which takes as input positive (e.g., regulated transcripts) and negative gene lists (e.g., the co-expressed transcripts not affected by the same post-transcriptional event), and outputs the RNA cis-regulatory element that distinguishes the positive from the negative transcripts. First, RNA-READ tests for enrichment of previously reported RNA motifs, then it performs novel RNA motif discovery to identify the consensus RNA motif that best discriminates between the positive and negative transcripts. An important innovation of the RNA-READ pipeline is that it considers both sequence and structural constraints on binding of RBPs. I consider two binding classes: (1) the RBP binds to mRNAs with primary sequence-specificities; (2) the RBP binds to elements in mRNAs defined completely by structure (i.e., shape recognition). I have shown that computationally estimated target-site accessibility improves prediction of sequence-specific binding for various RBPs, with >22% average relative decrease in error versus using only sequence information. The predictive power is further increased with the introduction of structural-context constraints to the single-stranded target sites. Furthermore, I showed that computationally estimated intrinsic mRNA secondary structure is also helpful for determining RBP binding via shape recognition. I identified specific structural elements enriched in dsRBP Staufen targets versus non-targets. I applied the RNA-READ
pipeline to the datasets that measured translation, stability and localization of transcripts in the
Drosophila early embryo and identified numerous significant associations between the presence
of motif matches and specific regulatory outcomes.
Acknowledgments

I would like to express my deepest appreciation to my mentors Dr. Howard Lipshitz and Dr. Quaid Morris for their patient guidance, continuous support and warm encouragement. I enjoyed every moment of my time spent with them on science. They showed me how to think over, pursue, and accomplish a scientific problem. I hope I can someday be as smart, creative, and enthusiastic as they are. Additionally, they were very accessible and supportive. As a student without any computer science background, I used to doubt my ability to work in the computational-biology field. They offered me much help to overcome that doubt. Their support is not only scientific in nature. I am also extremely appreciative for their understanding towards my leave of absence during my Ph.D. study, which allowed me to be with my father during his last year.

I would also like to thank my committee members, Dr. Henry Krause and Dr. Andrew Spence for giving insightful comments and suggestions on my projects. I appreciate their contributions of time, ideas, and encouragement that made my Ph.D. experience productive and stimulating. Additionally, I am extremely appreciative to have been supervised by Henry in the early years of my Ph.D study.

I have been fortunate enough to know the fun people in Dr. Morris’ lab and Dr. Lipshitz’s lab. We have shared many ideas scientific and nonscientific alike. I would like to send my special thanks to Gerald Quon, Sepand Mavandadi and Rachel Bevan, who helped me significantly with the computational field in my early Ph.D. years. In my wet-lab life, I would like to thank Najeeb Siddiqui, Hua Luo, Linan Emily Chen and Liang Ming for showing me how to do biology experiments, including RNA in situ hybridization, microarray, real-time PCR, and construct synthesis. I am also very appreciative to Angelo Karaiskakis for his great help on ordering experimental materials for me. In addition, I enjoyed the Morris’event hosted by Amit Deshwar, Starbucks coffee with Angelo and Zhiyong, and all the other social events with everyone in the lab. Just as I had been told, the co-supervision doubled my joy and social life in graduate school.

I am also very grateful to the collaborators I have interacted with during my Ph.D. I worked with Gerald Quon on evaluating the role of accessibility in RBP binding; with Najeeb Siddiqui and
Hua Luo on the study of primordial germ cell; with John Laver and Dr. Craig Smibert as well as Kristin Ancevicius and Dr. Tim Westwood on the Study of Staufen binding; with Hilal Kazan, Wei Jiao, Heng Pan and many people from Dr. Tim Hughes’s lab on the identification of RNA motifs. This thesis would not have been possible without their help.

I would like to thank my friends and family for their support. To Cara Liu, for always being there when I struggled with my life or studies. To the Zhou family, for their love, understanding and encouragement. To my grandparents and great-grandmother, for everything they taught me when I was a child. To my aunt Xiaoxia Li, for inspiring my initial interest in biology and continually encouraging me to pursue my career as a scientist.

I would like to especially thank my parents, Gang Li and Jianxin Liu, for being a constant source of love, support and encouragement during every stage of my life. They gave me life, gave me love and made me the person I am today. I deeply miss my father, Gang Li, who is not here today to share this joy with me. To him this thesis is dedicated.

Lastly, I would like to thank my husband Gao Zhou for his love and faithful support.
# Table of Contents

Acknowledgments ........................................................................................................ iv

Table of Contents ......................................................................................................... vi

List of Tables .................................................................................................................. x

List of Figures ................................................................................................................ xi

List of Appendices ......................................................................................................... xiii

Chapter 1 ......................................................................................................................... 1

Introduction ..................................................................................................................... 1

1.1 Post-transcriptional regulation .............................................................................. 2

1.2 How RNA-binding proteins bind RNA ................................................................. 2

1.3 Experimental methods to detect RNA-protein interactions ............................... 6

1.4 Computational methods that use primary sequence to identify RBP target sites .... 11

1.5 Evidence that intrinsic mRNA secondary structure has an impact on RBP binding .... 17

1.6 Computational methods for prediction of mRNA structure ............................... 18

1.7 Experimental methods for prediction of mRNA structure ................................. 21

1.8 Combining experimental and computational methods ......................................... 23

1.9 RBP motif discovery algorithms that incorporate secondary structure information .... 24

   1.9.1 Structural context-based methods .................................................................. 24

   1.9.2 Stochastic context-free grammar (SCFG)-based methods ............................ 26

1.10 Using sequence and structure conservation to find RBP binding sites ............ 27

1.11 Overview of thesis research ................................................................................. 28

Chapter 2 ....................................................................................................................... 30

Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure .... 30

2.1 Abstract .................................................................................................................... 31

2.2 Introduction ............................................................................................................. 31
2.3 Results........................................................................................................................................34
  2.3.1 Target site accessibility predicts mRNA targets of Pumilio and Puf3p.................. 34
  2.3.2 Accessibility improves motif-based mRNA target prediction for diverse RBPs.. 40
  2.3.3 Improvement due to accessibility is not explained by nucleotide composition
      biases........................................................................................................................................45
  2.3.4 Many RBPs require the entire target site to be accessible ......................... 49
  2.3.5 The impact of other variations in the calculation of target site accessibility...... 53
  2.3.6 In vivo sequence motif finding using accessibility ........................................... 53

2.4 Discussion...................................................................................................................................58

2.5 Methods.......................................................................................................................................61
  2.5.1 RBP co-purification and sequence motif data collection................................. 61
  2.5.2 Defining bound and unbound sets of transcripts........................................... 62
  2.5.3 Quantifying target site accessibility ................................................................. 62
  2.5.4 Scoring accessibility of target sites and their flanking regions....................... 63
  2.5.5 Statistical tests for the significance of difference between two AUROCs ......... 63
  2.5.6 Motif finding procedure and cross-validation.................................................. 64

2.6 Additional Supporting Data found in the accompanying CD ............................... 65

Chapter 3 .........................................................................................................................................66

Genome-wide analysis of Staufen-associated mRNAs identifies motifs that confer target
specificity ........................................................................................................................................66

3.1 Abstract.....................................................................................................................................67

3.2 Introduction...............................................................................................................................67

3.3 Results.......................................................................................................................................69
  3.3.1 Genome-wide identification of Staufen-associated mRNAs ............................ 69
  3.3.2 Drosophila and human Staufen targets have unusually long 3’UTRs............ 84
  3.3.3 High-confidence Drosophila Staufen target mRNAs are enriched for specific
      structural motifs in their 3’UTRs ...................................................................................... 88
3.3.4 Computational analysis of the properties of dsRNA stems bound by Staufen ..... 99
3.3.5 Definition of Staufen-recognized structures (SRSs) ................................. 107
3.3.6 Mapping of SRSs in *Drosophila* and human mRNAs .............................. 116

3.4 Discussion ........................................................................................................ 124
3.4.1 A role for 3’UTR length in Staufen-mediated decay? .............................. 124
3.4.2 Structural motifs that predict Staufen binding .................................. 125
3.4.3 How does Staufen recognize and bind stems of different lengths? ....... 126
3.4.4 Staufen levels as a determinant of target mRNA selection .................. 127

3.5 Materials and Methods .................................................................................. 127
3.5.1 *Drosophila* stocks ................................................................................. 128
3.5.2 RNA co-immunoprecipitations ............................................................. 128
3.5.3 Microarrays ............................................................................................. 128
3.5.4 Data access ............................................................................................. 129
3.5.5 Reverse transcription-quantitative PCR .............................................. 129
3.5.6 Source of transcript sequences for assessment of UTR and ORF lengths and motif finding ............................................................ 129
3.5.7 Definitions of secondary structure terms ............................................. 130
3.5.8 Defining *N of M* motif hits ................................................................. 130
3.5.9 Discovery of *N of M* motifs that predict Staufen binding ................... 131
3.5.10 Defining [12,10] and [19,15] structures ............................................... 132
3.5.11 Identification of additional features of Staufen-recognized [12,10] and [19,15] structures ............................................................... 132
3.5.12 Using Staufen-recognized structures (SRSs) to predict Staufen targets and non-targets .......................................................... 133
3.5.13 Scoring of the precision of motif mapping .......................................... 134
3.5.14 Defining bound and unbound sets for the human Staufens .................. 134

3.6 Additional Supporting Data found in the accompanying CD ..................... 135
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 4</strong></td>
<td>136</td>
</tr>
<tr>
<td>Development and application of the RNA-READ pipeline</td>
<td>136</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>137</td>
</tr>
<tr>
<td>4.2 Results</td>
<td>138</td>
</tr>
<tr>
<td>4.2.1 Comparison of experimental and computational methods for prediction of mRNA structure</td>
<td>138</td>
</tr>
<tr>
<td>4.2.2 Consideration of specific structural context further improves mRNA target prediction</td>
<td>142</td>
</tr>
<tr>
<td>4.2.3 RNA Regulatory Elements Analysis and Discovery (RNA-READ)</td>
<td>144</td>
</tr>
<tr>
<td>4.3 Discussion</td>
<td>150</td>
</tr>
<tr>
<td>4.4 Methods</td>
<td>154</td>
</tr>
<tr>
<td>4.4.1 Quantifying target site accessibility with specific structural restrictions</td>
<td>154</td>
</tr>
<tr>
<td>4.4.2 PTR categories</td>
<td>154</td>
</tr>
<tr>
<td>4.4.3 de novo motif discovery</td>
<td>156</td>
</tr>
<tr>
<td>4.4.4 Motif scanning</td>
<td>157</td>
</tr>
<tr>
<td>4.5 Additional Supporting Data found in the accompanying CD</td>
<td>159</td>
</tr>
<tr>
<td><strong>Chapter 5</strong></td>
<td>160</td>
</tr>
<tr>
<td>Summary and future directions</td>
<td>160</td>
</tr>
<tr>
<td>5.1 Summary of my work</td>
<td>161</td>
</tr>
<tr>
<td>5.2 Future direction</td>
<td>164</td>
</tr>
<tr>
<td>5.2.1 Gapped motif finders for RBPs</td>
<td>164</td>
</tr>
<tr>
<td>5.2.2 Methods to predict RNA tertiary structure</td>
<td>167</td>
</tr>
<tr>
<td>5.2.3 Combinatorial interactions among RBPs, miRNAs and mRNAs</td>
<td>167</td>
</tr>
<tr>
<td>5.2.4 Experimental strategies for testing putative RNA cis-regulatory element</td>
<td>169</td>
</tr>
<tr>
<td>References</td>
<td>171</td>
</tr>
</tbody>
</table>
List of Tables

TABLE 1.1. WEB RESOURCES FOR RBP BINDING SITES ................................................................. 10
TABLE 1.2. MOTIF FINDING METHODS ......................................................................................... 16
TABLE 1.3. WEB RESOURCE FOR PREDICTING MRNA SECONDARY STRUCTURE ....................... 20
TABLE 2.1: TARGET SITE ACCESSIBILITY PREDICTS MRNA TARGETS OF PUMILO AND PUF3P .... 39
TABLE 2.2. THE IMPACT OF OTHER VARIATIONS IN THE CALCULATION OF TARGET SITE ACCESSIBILITY ................................................................. 52
SUPPLEMENTARY FILE S2.1. DETAILS OF CO-PURIFICATION DATASETS; PREVIOUSLY DEFINED RBP

  CONSENSUS SEQUENCES; DETAILED DESCRIPTION OF ANALYZED RBPs; AND MOTIF FINDING

  RESULTS FROM 3X10 CROSS-VALIDATION TESTS. ....................................................................... 65
SUPPLEMENTARY FILE S2.2. EXACT P-VALUES FOR ALL ANALYSES............................................. 65
TABLE 3.1. STAUFEN-ASSOCIATED MRNAS IDENTIFIED BY SYNTHETIC ANTI-STAUFEN RIP-CHIP (FOLD-

  ENRICHMENT ≥ 2) AND ANTI-GFP-STAUFEN RIP-CHIP (FOLD-ENRICHMENT ≥ 5) ...................... 73
TABLE 3.2. SEARCH RESULTS OF THE SIX COMPUTATIONALLY IDENTIFIED MOTIFS IN DROSOPHILA 5'UTRS,

  ORFS AND 3'UTRS......................................................................................................................... 98
TABLE 3.3. SFOLD VALIDATION OF 10 OF 12 AND 15 OF 19 MOTIF HITS. .................................... 101
SUPPLEMENTARY TABLE S3.1. LIST OF MRNA TARGETS IDENTIFIED BY SYNTHETIC ANTI-STAUFEN RIP-

  CHIP ........................................................................................................................................... 135
SUPPLEMENTARY TABLE S3.2. RT-QPCR VERIFICATION OF MICROARRAY RESULTS FOR SELECTED MRNAS.

  .................................................................................................................................................. 135
SUPPLEMENTARY TABLE S3.3. LIST OF MRNA TARGETS IDENTIFIED BY ANTI-GFP RIP-CHIP ........ 135
SUPPLEMENTARY TABLE S3.4. LENGTH COMPARISONS BETWEEN THE RBP TARGETS AND CO-EXPRESSED

  NON-TARGETS PRESENTED IN FIGURE 4 AND SUPPLEMENTARY FIGURE S6 .............................. 135
SUPPLEMENTARY TABLE S3.5. CUTOFF OF TOP 1% OF THE N OF M PROBABILITIES ACROSS ALL 3'UTRS.135
SUPPLEMENTARY TABLE S3.6. DETAILS OF THE DOUBLE-STRANDED MOTIF SEARCHES IN DROSOPHILA

  3'UTRS ....................................................................................................................................... 135
SUPPLEMENTARY TABLES S3.7. LOCATION OF THE PREDICTED SRSS IN THE 3'UTRS OF ALL DROSOPHILA

  MRNAS ...................................................................................................................................... 135
SUPPLEMENTARY TABLE S4.1. RESULTS OF RNA-READ ON POST-TRANSCRIPTIONAL REGULATION

  CATEGORIES .............................................................................................................................. 159
SUPPLEMENTARY TABLE S4.2. IUPAC MOTIFS DEFINED BY RNACOMPETE ................................... 159
List of Figures

FIGURE 1.1. COMPARISON OF DSDNA AND DSRNA HELIX.............................................................................. 4
FIGURE 1.2. THREE-DIMENSIONAL STRUCTURES OF RBD-RNA COMPLEXES......................................................... 5
FIGURE 2.1. PUF3P AND PUMILIO CONSENSUS BINDING SITES HAVE HIGHER ACCESSIBILITY IN 3’UTRS
OF THEIR BOUND MRNA TARGETS.......................................................................................................................... 37
FIGURE 2.2. PUM1 SEQUENCE CONSENSUS SITE HAVE HIGHER ACCESSIBILITY IN 3’UTRS OF THEIR
MRNA TARGETS......................................................................................................................................................... 38
FIGURE 2.3. SCHEMATIC OF THE IN SILECO ASSAY FOR MEASURING THE IMPACT OF TARGET SITE
ACCESSIBILITY ON RBP BINDING............................................................................................................................. 42
FIGURE 2.4. TARGET SITE ACCESSIBILITY PREDICTS IN VIVO BINDING FOR A DIVERSE RANGE OF RBPS.
..................................................................................................................................................................................... 44
FIGURE 2.5. 3’UTR TARGET SITE ACCESSIBILITY PREDICTS IN VIVO BINDING............................................. 46
FIGURE 2.6. DIFFERENCES IN DINUCLEOTIDE COMPOSITION AROUND PUTATIVE RBP BINDING SITES
BETWEEN BOUND AND UNBOUND TRANSCRIPTS................................................................................................. 47
FIGURE 2.7. COMPARATIVE PREDICTIVE ACCURACY OF #TS AND #ATS WHEN REVERSE CONSENSUS
MOTIFS ARE SCORED..................................................................................................................................................... 48
FIGURE 2.8. TARGET SITE ACCESSIBILITY IS A BETTER PREDICTOR THAN AVERAGE/MINIMAL
ACCESSIBILITY OF SINGLE BASES IN THE TARGET SITE ..................................................................................... 51
FIGURE 2.9. RBP MOTIFS OPTIMIZED TO DISTINGUISH BOUND VERSUS UNBOUND TRANSCRIPTS........... 56
FIGURE 2.10. LEARNED IN VIVO MOTIFS FOR SIX YEAST RBPS........................................................................ 57
FIGURE 3.1. ENRICHMENT OF EXPRESSED TRANSCRIPTS IN STAUFEN RIPS USING SYNTHETIC ANTI-
STAUFEN AND ANTI-GFP STAUFEN ANTIBODIES............................................................................................... 71
FIGURE 3.2. MICROARRAY EXPRESSION OF TRANSCRIPTS FROM W1118 EMBRYO EXTRACT USED AS
INPUT FOR SYNTHETIC ANTI-SHAUFEN RIPS........................................................................................................ 74
FIGURE 3.3. REPLICATE-TO-REPLICATE COMPARISONS OF TRANSCRIPT MICROARRAY SIGNAL
INTENSITIES FROM THE IMMUNOPRECIPITATED SAMPLES OF THE SYNTHETIC ANTI-SHAUFEN
AND CONTROL SYNTHETIC ANTIBODY RIPS...................................................................................................... 75
FIGURE 3.4. WESTERN BLOTS EXAMINING STAUFEN IMMUNOPRECIPITATION RECOVERY AND
STAUFEN EXPRESSION LEVELS.............................................................................................................................. 77
FIGURE 3.5. MICROARRAY EXPRESSION OF TRANSCRIPTS FROM GFP-SHAUFEN EMBRYO EXTRACT
USED AS INPUT FOR ANTI-GFP-SHAUFEN RIPS.................................................................................................... 78
FIGURE 3.6. REPLICATE-TO-REPLICATE COMPARISONS OF TRANSCRIPT MICROARRAY SIGNAL
INTENSITIES FROM THE SAMPLES OF THE ANTI-GFP AND CONTROL ANTI-FLAG RIPS.............................. 79
FIGURE 3.7. COMPARISON OF THE SYNTHETIC ANTI-SHAUFEN AND ANTI-GFP-SHAUFEN RIPS............ 83
FIGURE 3.8. DROSOPHILA AND HUMAN STAUFEN TARGETS HAVE UNUSUALLY LONG 3’UTRS. .............. 86
FIGURE 3.9. SINGLE-STRANDED RBP TARGETS HAVE LONG 3’UTRS............................................................ 87
FIGURE 3.10. SCHEMATIC OF THE IN SILICO ASSAYS FOR DISCOVERY OF STAUFEN’S BINDING PREFERENCES. .................................................................................................................................................. 89

FIGURE 3.11. SPECIFIC DOUBLE-STRANDED STRUCTURES ARE ENRICHED IN STAUFEN TARGET TRANSCRIPT 3’UTRs. ................................................................................................................................................. 93

FIGURE 3.12. CORRELATIONS BETWEEN THE SCORES OF THE DOUBLE-STRANDED MOTIFS SEARCHED FOR IN DROSOPHILA 3’UTRS. ..................................................................................................................... 95

FIGURE 3.13. CHARACTERISTICS OF THE STEMS BOUND BY STAUFEN AT THE LEVEL OF INDIVIDUAL SITES. ........................................................................................................................................................................... 103

FIGURE 3.14. LOOP SIZE IN POSITIVE AND NEGATIVE SETS......................................................................................................................... 104

FIGURE 3.15. CHARACTERISTICS OF THE STEMS BOUND BY STAUFEN. ........................................................................................................ 106

FIGURE 3.17. CLASSIFICATION OF TRANSCRIPTS BASED ON SRS TYPE...................................................................................................... 114

FIGURE 3.18. DISTANCE BETWEEN THE ARMS OF DSrna STEMS.................................................................................................................. 115

FIGURE 3.19. MAPPING OF SRSS IN THE 3’UTRS OF STAUFEN TARGETS AND NON-TARGETS. .............................................................. 120

FIGURE 3.20. PREDICTED STAUFEN-BINDING MOTIFS MAP WITH HIGH PRECISION TO THE KNOWN STAUFEN-BINDING REGIONS IN DROSOPHILA BICOID AND HUMAN ARFI 3’UTRS................................................. 123

FIGURE 4.1. COMPARISON OF PREDICTION ACCURACY FOR IN VIVO BINDING OF NINE YEAST RBPS USING PARS AND Rnaplfold TO ESTIMATE THE SECONDARY STRUCTURE OF BOUND VERSUS UNBOUND TRANSCRIPTS. ........................................................................................................................................... 141

FIGURE 4.2. STRUCTURAL CONTEXT OF TARGET SITES IMPROVES PREDICTION OF TARGET MRNAS BOUND IN VIVO BY RBPS. ........................................................................................................................................... 143

FIGURE 4.3. FLOWCHART OF RNA-READ PIPELINE. ................................................................................................................................. 144

FIGURE 4.4. SMAUG RECOGNITION ELEMENTS (SRES) ARE ENRICHED IN SMAUG ASSOCIATED TARGETS AND SMAUG REGULATED TARGETS IN DROSOPHILA. ................................................................................................. 146

FIGURE 4.5. PRELIMINARY RNA-READ MOTIF DISCOVERY RESULTS ON FLY-FISH DATA. ........................................................................ 148

FIGURE 4.6 FUTURE DIRECTION OF RNA-READ. ........................................................................................................................................... 153

FIGURE 5.1. TARGET SITE ACCESSIBILITY PREDICTS IN VIVO BINDING FOR A DIVERSE RANGE OF RBPS. ........................................................................................................................................................................... 162

FIGURE 5.2. THREE-DIMENSIONAL STRUCTURES OF MULTIPLE RBDS IN COMPLEX WITH RNA. ................................................................. 166
List of Appendices

Acknowledgments ................................................................................................................ iv
Table of Contents ................................................................................................................... vi
List of Tables .......................................................................................................................... x
List of Figures ....................................................................................................................... xi
List of Appendices ............................................................................................................... xiii
References ............................................................................................................................ 171
Chapter 1

Introduction

Some sections of this chapter are derived from the following article: Li, X.*, Kazan, H.*, Lipshitz, H.D., Morris, Q., Finding the target sites of RNA-binding proteins. WIRE, submitted. I wrote all the content in this Chapter except Sections 1.4, 1.6 and 1.9, which were drafted by Hilal Kazan and updated by me afterwards.
1.1 Post-transcriptional regulation

Post-transcriptional regulation plays a critical role in the control of gene expression during many biological processes, including the cell cycle, development, inflammatory and immune responses\textsuperscript{1-3}. Post-transcriptional regulation controls gene expression at different levels, including translational repression\textsuperscript{4}, RNA stability\textsuperscript{5}, and RNA localization\textsuperscript{6}. These regulatory processes are often mediated by the binding of RNA binding proteins (RBPs) or small RNAs to the \textit{cis}-regulatory elements in the transcripts\textsuperscript{2,7}. Identification of RNA \textit{cis}-regulatory element is the first step towards understanding the mechanisms by which they conduct post-transcriptional regulation.

Eukaryotic genomes encode hundreds of RNA-binding proteins (RBPs) with diverse functions in co- and post-transcriptional regulation of RNA metabolism. Recent studies have revealed that RBPs typically have hundreds of targets and multiple RBPs coordinately regulate populations of functionally related mRNAs\textsuperscript{8-11}. Identification of RBP target sites is an important step towards understanding the mechanisms by which they conduct post-transcriptional regulation.

1.2 How RNA-binding proteins bind RNA

Primary sequence specificity is often critical for binding-site recognition by both RNA- and DNA-binding proteins; however, RNA-protein interactions differ from DNA-protein interactions because double-stranded RNA (dsRNA) typically adopts the A-form helical structure whose major groove is deeper and narrower than that of the B-form helix of dsDNA (Figure 1.1). As such, base-specific interactions by amino acid side chains are rare in dsRNA\textsuperscript{12,13} and sequence-specific RBPs are likely to require at least some of their binding site to be single-stranded\textsuperscript{14}. There is substantial evidence that this is the almost exclusive form of interaction between sequence-specific RBPs and their targets. Indeed, the two most common RNA-binding domains
in eukaryotes, the RNA Recognition Motif (RRM) and the hnRNP K-homology (KH) domains bind single-stranded RNA\textsuperscript{15,16}. Comprehensive surveys\textsuperscript{17-19} of RBP-RNA complexes deposited in the Protein Data Bank (PDB) have reported that base-specific interactions between RBPs and RNA only occur in or near regions of single-stranded RNA (ssRNA). As examples of RBP interactions with ssRNA, Figures 1.2A and 1.2B show structures of an RRM and Pum-homology domains (PUM-HD) in complex with their ssRNA targets.
Figure 1.1. Comparison of dsDNA and dsRNA helix.
Figure 1.2. Three-dimensional structures of RBD-RNA complexes.

(A) Solution structure of Polypyrimidine Tract Binding (PTB) protein RBD1 in complex with CUCUCU RNA (PDB: 2AD9). PTB RBD1 binds a YCU site (Y indicating pyrimidine) through ß4, ß1 and ß2, respectively. (B) Co-crystal structure of the PUM-homology domain (PUM-HD) in human Pum1 complexed with a 10-nucleotide single-stranded RNA, 5'-AUUGUACAUA where the last eight nucleotides (UGUACAUA) are individually recognized by three conserved amino acids in Puf repeats 8-to-1, respectively (PDB: 1M8Y). (C) Solution structure of the Vts1p sterile-α motif (SAM) domain in complex with a 5'-CUGGC-3' pentaloop as part of a 19nt hairpin (PDB: 2ESE). The specific interaction between the Vts1p SAM domain and the target RNA is stabilized by both the direct interaction to the third guanosine base in the RNA pentaloop and the contacts to the unique backbone structure. (D) Solution structure of dsRBD of yeast Rnt1p in complex with the 5’ terminal AGNN tetraloop of snR47 precursor RNA (PDB: 1T4I). Neither A nor G are recognized by specific hydrogen bonds; instead, the N-terminal helix of the Rnt1p dsRBD interacts with the backbone and the two nonconserved tetraloop bases, by snugly fitting into the minor groove side of the RNA tetraloop and extending into the minor groove at the top of the stem.
1.3 Experimental methods to detect RNA-protein interactions

Identification of the RNAs bound by each RBP is key for understanding the interactions governing post-transcriptional regulation. A number of low- and high-throughput experimental methods have been developed to assess the in vitro sequence-binding preferences of RBPs, as well as to identify the in vivo binding sites for RBPs in particular cellular contexts (see Table 1.1 for the web resources for RBP binding sites).

SELEX (Systematic Evolution of Ligands by Exponential Enrichment) is a low-throughput method for in vitro detection of RBP sequence-binding preferences\textsuperscript{25}. High-affinity binding sequences are selected from a randomized RNA oligonucleotide pool by several sequential rounds of binding to purified protein, each followed by PCR amplification. The products are then cloned and sequenced, identifying a set of short sequences preferred by the protein. These short sequences are then analyzed in order to define primary sequence and structural preferences of the RBP. One disadvantage of the SELEX assay is that, because of the multiple rounds of purification and amplification, it reveals only the highest affinity RNA target sites, and does not completely characterize the range, and relative affinity of RNA-sequence preferences of an RBP.

The recent advent of relatively inexpensive, high-throughput sequencing has facilitated the development of a more quantitative and comprehensive version of this procedure, sometimes called HT-SELEX\textsuperscript{26,27}. In this procedure, only a single, or a small number, of binding reactions is performed but millions of RNA oligos are sequenced, supporting a more quantitative estimate of the RBP sequence-binding preference.

RNAcompete is a related in vitro method that replaces the large, complex random initial RNA oligo pool used by HT-SELEX with a smaller, carefully designed pool that is synthesized with the help of a custom designed microarray. The oligo pool contains approximately 244,000 short
30-38nt RNAs whose design is based on modified de Bruijn sequences\textsuperscript{16,28,29}, that ensure that 7nt RNA sequences appear either in ssRNA or weakly-paired RNA in at least 64 oligos. This allows an unbiased measurement of the relative sequence-binding preferences of RBPs. An advantage of RNAcompete is that it is much less expensive than HT-SELEX because its small pool size allows the relative abundances of each oligo to be measured using a custom-designed Agilent microarray. To date, RNA primary sequence preference for more than 200 RBPs have been reported and these are summarized in the cisbp-RNA website\textsuperscript{30,31}. However, because the RNAcompete pool is depleted for RNAs with stable secondary structure, RBPs with strict structural requirements on their binding sites are less successful in the assay. Nonetheless, RNAcompete is still able to recover the primary sequence binding preferences of some RBPs that have preferences for particular secondary structural contexts, such as Vts1p\textsuperscript{21-23} and Lin28\textsuperscript{32}.

There are two major approaches for large-scale assays of RBP binding sites \textit{in vivo}: Ribonucleoprotein Immunoprecipitation (RIP) based methods, which do not cross-link the RBP to the RNA, and Cross-Linking and Immunoprecipitation (CLIP) based methods, which do. In RIP-based assays, RNAs associated with the RBP of interest are isolated from cell lysate after immunoprecipitation of the RBP, and then identified using either microarray or sequencing technologies\textsuperscript{33}. CLIP-based assays use ultraviolet (UV) light to form cross-links between RNAs and the RBP, followed by use of ribonuclease to partially digest the bound RNAs\textsuperscript{34}, leaving only small segments that are in direct contact with the RBP. Although RIP-based assays are simpler and more widely applicable, because UV-based crosslinking is difficult in some cells or tissues\textsuperscript{35}, the irreversible covalent bond introduced by cross-linking allows a more stringent washing procedure in CLIP, which reduces the number of false-positive targets during the purification step. Cross-linking also protects the target site from ribonuclease digestion, allowing a much greater resolution in determining the actual site of interaction. One method, Photoactivatable-Ribonucleoside–Enhanced Cross-linking and Immunoprecipitation (PAR-CLIP) modifies CLIP by culturing living cells with a photoreactive ribonucleoside analogue, such as 4-thiouridine (4-
SU), to facilitate cross-linking\textsuperscript{36}. The chemical structural change of the 4-SU base upon cross-linking to an the RBP causes preferential pairing of guanine (G) rather than adenine (A) with the 4-SU base, and therefore introduces a thymine (T) to cytosine (C) transition at the cross-linked position during PCR amplification. In PAR-CLIP, the frequencies and types of mutations observed are used as indicators to pinpoint the precise RBP binding site\textsuperscript{37}. Diagnostic mutations are also observed in other CLIP approaches, though with a lower frequency\textsuperscript{38}. Although these techniques increase the resolution of these methods, they still cannot robustly achieve single-nucleotide resolution\textsuperscript{38}. 
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Function</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ARESITE</strong></td>
<td>AU-rich elements (ARE) in vertebrate mRNA UTR sequences</td>
<td>Input gene sequence is searched for enrichment of eight predefined consensus ARE. For each detected motif, conservation patterns and predicted accessibility values are displayed.</td>
<td><a href="http://rna.tbi.univie.ac.at/AREsite/">http://rna.tbi.univie.ac.at/AREsite/</a></td>
</tr>
<tr>
<td><strong>CisBP-RNA</strong></td>
<td>RBP binding sites identified by RNAcompete</td>
<td>Users can search or browse RBP binding sites identified by RNAcompete method.</td>
<td><a href="http://cisbp-rna.ccbr.utoronto.ca/">http://cisbp-rna.ccbr.utoronto.ca/</a></td>
</tr>
<tr>
<td><strong>CLIPZ</strong></td>
<td>Binding sites from CLIP experiments, including Quaking, Pumilio, Argonautes 1-4, TNRC6 A-C, IGF2BP 1-3</td>
<td>Users can browse the clusters of genome- or transcript-based reads. Clusters from different experiments can be compared. The transcripts associated with a gene name could be searched for binding sites. There is also a motif enrichment tool that identifies over-represented nmers in a set of sequences.</td>
<td><a href="http://www.clipz.unibas.ch/">http://www.clipz.unibas.ch/</a></td>
</tr>
<tr>
<td><strong>doRiNA</strong></td>
<td>RBP and miRNA binding sites identified by CLIP experiments</td>
<td>CLIP-derived peaks for RBPs and miRNAs from humans, mouse, flies, and worms are available. Users can also search overlapping sites between multiple RBPs or between RBPs and miRNAs.</td>
<td><a href="http://dorina.mdc-berlin.de/">http://dorina.mdc-berlin.de/</a></td>
</tr>
<tr>
<td><strong>RBPDB</strong></td>
<td>Experiments and observations about RBP binding sites in metazoan genomes</td>
<td>All experiments related to an RBP could be retrieved by entering the associated gene name. Input sequences could be scanned for matches with RBP binding sites.</td>
<td><a href="http://rbpdb.ccbr.utoronto.ca/">http://rbpdb.ccbr.utoronto.ca/</a></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Rfam</strong></td>
<td>Non-coding RNA genes, structured cis-regulatory elements and self-splicing elements</td>
<td>Each entry includes multiple sequence alignment, a secondary structure and related references. Please see [citation] for a complete description of available features.</td>
<td><a href="http://rfam.sanger.ac.uk/">http://rfam.sanger.ac.uk/</a></td>
</tr>
<tr>
<td><strong>UTRSite</strong></td>
<td>Regulatory elements in 5’ and 3’ UTRs</td>
<td>Each entry summarizes the current knowledge on a regulatory element: location (e.g. 3’UTR), Rfam cross-reference, binding proteins and interactor(s) of binding protein(s) and related references. Tools for searching and scanning are available.</td>
<td><a href="http://utrsite.ba.itb.cnr.it/">http://utrsite.ba.itb.cnr.it/</a></td>
</tr>
</tbody>
</table>

**Table 1.1. Web Resources for RBP binding sites**
1.4 Computational methods that use primary sequence to identify RBP target sites

Even if experimentally defined RBP-binding sites are available, computational motif-based methods are useful to define the precise site of binding, to detect false positives and negatives, to identify degenerate motifs, to model the impact of RNA secondary structure on binding, to identify co-binding factors (e.g.,\textsuperscript{45}), and to predict the likely impact of polymorphisms on RBP-RNA interactions.

Often motif models developed for DNA-binding proteins have been adapted to identify primary sequence preferences of RBPs and to scan transcripts for potential binding sites. For example, MatrixREDUCE was used to find RNA motifs associated with transcript stability in yeast\textsuperscript{46} and to recover binding preferences of RBPs from in vitro binding affinity data\textsuperscript{47}. This model represents the binding sites with a position specific affinity matrix (PSAM) that can be used to predict the relative affinity for each potential binding site. Unlike many other motif-finding methods, MatrixREDUCE takes as input quantitative values associated with each sequence in the dataset rather than a subset pre-defined as ‘bound’ or ‘unbound’. MEME (Multiple Expectation Maximization for Motif Elicitation)\textsuperscript{48} is another popular motif discovery algorithm originally designed to find repeated, ungapped sequence patterns in DNA or proteins. MEME has been used to predict motifs for Puf proteins in flies and yeast\textsuperscript{49,50}. Additional models, such as FIRE (Finding Informative Regulatory Elements)\textsuperscript{51}, and REFINE (Relative Filtering by Nucleotide Enrichment)\textsuperscript{52}, have been used to identify a group of sequence consensuses from yeast RIP-Chip datasets\textsuperscript{8}. A summary of motif discovery tools can be found in Table 1.2. Users who choose to run DNA-motif finders on RNA should adjust the options within these methods (e.g., searching complementary strands should be turned off).
Primary sequence motif-based models can miss important secondary structural context constraints and, in doing so, incorrectly predict the primary sequence preference of the RBP\textsuperscript{53}. For example, both REFINE and FIRE fail to identify known binding preferences of Vts1p \textit{(i.e., CNGG within a hairpin loop)} from RIP-Chip data\textsuperscript{8}, whereas this primary sequence motif is easily found on the same data by motif finders that also model preferences for RNA accessibility\textsuperscript{54}. In the following sections, I review the evidence to support a role for mRNA secondary structure in sequence-specific RBP binding and I then describe methods used to determine mRNA secondary structure.
<table>
<thead>
<tr>
<th>Software / method</th>
<th>Input</th>
<th>Summary</th>
<th>Availability</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aptamotif</strong></td>
<td>RNA sequences identified by SELEX</td>
<td>A method for finding sequence-structure motifs in SELEX-derived aptamers. RNA secondary structure is predicted with ensemble-based methods.</td>
<td>Software package: available upon request</td>
<td>55</td>
</tr>
<tr>
<td><strong>cERMIT</strong></td>
<td>DNA or RNA sequences and associated expression or affinity measures</td>
<td>A rank-ordered based method that searches for sequence motifs bested supported by the observed experimental evidence (i.e., quantitative genome-wide binding data). It uses the complete dataset and does not require a cutoff to define the positive set.</td>
<td>Software package: <a href="http://www.genome.duke.edu/labs/ohler/research/transcription/cERMIT/">http://www.genome.duke.edu/labs/ohler/research/transcription/cERMIT/</a></td>
<td>57</td>
</tr>
<tr>
<td>Software</td>
<td>Description</td>
<td>Web server</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FIRE</strong></td>
<td>A method to detect DNA or RNA motifs that model the mutual information between sequences and gene expression measurements.</td>
<td><a href="http://tavazoie.lab.princeton.edu/FIRE/">http://tavazoie.lab.princeton.edu/FIRE/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MatrixREDUCE</strong></td>
<td>A biophysical model to discover sequence-specific binding affinity of the factor of interest (TF or RBP)</td>
<td><a href="http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/">http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MEME</strong></td>
<td>A generative model for finding motifs in DNA or protein sequences. Can be used for finding sequence motifs in RNA sequences.</td>
<td><a href="http://meme.sdsu.edu/meme4_6_0/cgi/meme.cgi">http://meme.sdsu.edu/meme4_6_0/cgi/meme.cgi</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tool</td>
<td>Input</td>
<td>Description</td>
<td>Software</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>MEMERIS</td>
<td>RNA sequences and predicted structures</td>
<td>Extension of MEME for finding RNA motifs. It uses RNA structure information as a prior of motif start to guide the motif search towards single-stranded regions.</td>
<td><a href="http://www.bioinf.uni-freiburg.de/~hiller/MEMERIS/">http://www.bioinf.uni-freiburg.de/~hiller/MEMERIS/</a></td>
<td></td>
</tr>
<tr>
<td>REFINE</td>
<td>DNA or RNA sequences</td>
<td>Extension of MEME, filters out regions of target sequences that are relatively devoid of discriminatory hexamers, and then applies MEME motif-finding algorithm.</td>
<td><a href="http://nar.oxfordjournals.org/content/early/2010/10/18/nar.gkq920/suppl/DC1">http://nar.oxfordjournals.org/content/early/2010/10/18/nar.gkq920/suppl/DC1</a></td>
<td></td>
</tr>
<tr>
<td><strong>RNAcontext</strong></td>
<td>RNA sequences, associated affinity measures and predicted structures</td>
<td>A discriminatory approach for finding RNA motifs that represent the sequence and structure preferences of RBPs. RNAcontext can model a wide range of structure features using a flexible alphabet.</td>
<td>Software package (includes scripts for structure prediction) <a href="http://morrislab.med.utoronto.ca/software.html">http://morrislab.med.utoronto.ca/software.html</a></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>RNApromo</strong></td>
<td>RNA sequences</td>
<td>CM-based model for finding RNA motifs</td>
<td>Software package and web server: <a href="http://genie.weizmann.ac.il/pubs/ramotifs08/rnamotifs08_predict.html">http://genie.weizmann.ac.il/pubs/ramotifs08/rnamotifs08_predict.html</a></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.2. Motif finding methods
1.5 Evidence that intrinsic mRNA secondary structure has an impact on RBP binding

The “accessibility” of a potential RBP target site plays an important role in whether the RBP actually binds to the site. Often this accessibility is defined based on predictions of mRNA secondary structure and can be roughly interpreted as the proportion of mRNA transcripts in which that site is single-stranded. This calculation is based exclusively on the RNA sequence without consideration of a potential role for other binding factors. The role of accessibility in binding-site selection by microRNAs (miRNAs) and small interfering RNAs (siRNAs) is well established\textsuperscript{61-64}; the role of accessibility in RBP-RNA interaction has taken longer to establish due to the diversity and complexity of these interactions.

The effect of RNA secondary structure in recognition of target sites was first investigated for the RBP, HuR\textsuperscript{65}. A positive correlation was found between predicted site accessibility and the binding affinity of HuR to sites that matched an NNUUNUUU HuR consensus. Furthermore, it was possible to alter HuR-mRNA binding in vitro and to increase mRNA stability in cell lysates by introducing secondary structure modulators that either increased or decreased the predicted accessibility of the HuR binding site; this was accomplished by hybridization to complementary RNAs that were predicted either to ‘open’ or ‘close’ the HuR binding site within the target mRNA’s secondary structure. Reduction in predicted accessibility also explained the reduction of HuR-TNFα binding upon an insertion of a sequence adjacent to the AU-rich element (ARE) in TNFα\textsuperscript{66,67}. A companion paper contained the “Hackermüller-Stadler” model, which models the observed Kd of an RBP to a bound RNA as the product of the probability (in the RNA structure ensemble) that the site is in the preferred structural context and the Kd of the RBP for the site in this context\textsuperscript{68}.
I have explored the role of accessibility in RBP-target interactions for more than a dozen RBPs from yeast, flies and humans; these RBPs contain a range of RBDs with diverse primary sequence binding preferences (described in detail in Chapter 2).

1.6 Computational methods for prediction of mRNA structure

The conclusion that the intrinsic mRNA secondary structure aids in finding RBP binding sites is plausible, but this also introduces a number of issues, most notably how to accurately determine mRNA secondary structure. By far the easiest way to assess mRNA secondary structure is using computational prediction methods (reviewed in this section, Table 1.3 for the web sources). However, the accuracy of these methods is controversial and, recently, biochemical methods have been introduced that query mRNA secondary structure genome-wide (see next section).

The most popular computational method to fold a single RNA sequence is based on the calculation of free energy from thermodynamic parameters derived from chemical melting experiments. Often, the focus is on the structure with the minimum free energy (MFE) because it is assumed that the RNA sequence folds into the lowest free energy structure at equilibrium. However, since thermodynamic parameters have substantial uncertainties and RNA secondary structure is often dynamic, the predicted MFE structure may not accurately represent the typical physical interactions that occur in the structure. To address these concerns, some methods consider the ensemble of all possible structures. One way to represent this ensemble is to use the centroid structure, which is defined as the structure with minimum total base-pair distance to all other structures in the ensemble.

Another way is to calculate base-pair probabilities from all possible structures using the partition function, with the assumption that the structure ensemble of any specific RNA sequence is Boltzmann distributed. However, accurately predicting the global structure of an mRNA is
challenging due to the decreasing predictive power of computational methods with increasing length of the input RNA\textsuperscript{81}. For long mRNA sequences, it is, in fact, often more accurate to only estimate structure using local interactions among bases and to ignore any potential long-range pairings\textsuperscript{82}. RNAplfold is one method for predicting site accessibility by averaging across short windows of the mRNA centered on the site of interest\textsuperscript{76,77}.

Other approaches predict RNA secondary structure based on pairwise covariation in multiple alignments with the assumption that functional RNA families should have conserved patterns of base pairing. Covariance models (CM) are a specialized stochastic context free grammar (SCFG) that probabilistically models both the RNA secondary structure and the primary sequence consensus of an RNA family\textsuperscript{58,83}. CMs are fit through a procedure that iterates between aligning individual sequences to a single CM and refining the CM based on the alignment\textsuperscript{58,83}. These methods work best when a good initial alignment is available to seed the search, and are used to predict families of functional RNAs (like tRNAs); however, their ability to model RBP binding sites in general is unclear. Indeed, the main challenge for predicting consensus structure from multiple sequences is that accurate structure prediction requires an accurate multiple-sequence alignment. Not only is simultaneously folding and aligning sequences computationally challenging\textsuperscript{84}, but this strategy may not be appropriate for modeling RBP-binding sites, as only the parts of the secondary structure that affect binding by the RBP may be conserved.
<table>
<thead>
<tr>
<th>Software / method</th>
<th>Input</th>
<th>Summary</th>
<th>Availability</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNAshapes</td>
<td>RNA sequence</td>
<td>It calculates shapes and their probabilities by analyzing the full ensemble, predicts the complete set of suboptimal structures and their probabilities</td>
<td>Software package and web server: <a href="http://bibiserv.techfak.unibielfeld.de/rnashapes/">http://bibiserv.techfak.unibielfeld.de/rnashapes/</a></td>
<td>85</td>
</tr>
<tr>
<td>RNAstructure</td>
<td>RNA sequence</td>
<td>It includes algorithms for RNA secondary structure prediction and calculation of base pair probabilities.</td>
<td>Software package with GUI: <a href="http://rna.urmc.rochester.edu/RNAstructure.html">http://rna.urmc.rochester.edu/RNAstructure.html</a></td>
<td>86</td>
</tr>
<tr>
<td>SFOLD</td>
<td>RNA sequence</td>
<td>It computes base pair probabilities from a representative sample of the full ensemble</td>
<td>Software package and web server: <a href="http://sfold.wadsworth.org/">http://sfold.wadsworth.org/</a></td>
<td>80</td>
</tr>
</tbody>
</table>

Table 1.3. Web resource for predicting mRNA secondary structure
1.7 Experimental methods for prediction of mRNA structure

Physical methods, including X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, have been used to describe RNA three-dimensional structure in great detail, but are often time-consuming and are limited to relatively short RNAs. RNA footprinting, an easier alternative, is often selected to analyze the structure of long RNAs. RNA footprinting detects RNA structure by treating the RNA of interest with a chemical or a nuclease to modify or cleave bases, respectively, that have a particular structural conformation (e.g. single-stranded, double-stranded or solvent-exposed)\(^{88-93}\). RNase-cleaved products, usually radioactively end-labeled, are then detected by autoradiography whereas the chemically modified bases are detected by electrophoresis of the reverse transcribed products that have stalled at the modified bases.

RNA footprinting has been extended to a large-scale method by combining next-generation sequencing technology with traditional RNase/chemical footprinting to simultaneously probe a mixture of RNAs\(^{94-96}\). Structural probing by chemical modification has a higher resolution than by nucleases because it is less restricted by steric hindrance; however, the read-out of the modification is much more difficult and, to date, has not been applicable to genome-wide assays of mRNA secondary structure. Selective 2’-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) is the major chemical modification-based technique. Unlike other base-selective chemical reagents, the hydroxyl-selective electrophiles used in SHAPE prevent reverse transcription on flexible nucleotides (e.g., single-stranded ones) by reacting with the 2’-hydroxyl group to form a 2’-O-adduct. This method can thereby interrogate all nucleotides in an RNA molecule simultaneously without biases towards certain primary sequences\(^{97}\). SHAPE-seq, which couples SHAPE chemistry with a multiplexed hierarchical barcoding and deep sequencing strategy\(^{96}\), has been used to accurately and simultaneously probe structures of several \textit{in vitro}
transcribed RNAs\textsuperscript{96}. However, the barcodes must be designed to target specific RNAs, thus preventing the expansion of SHAPE-seq to genome-wide assays\textsuperscript{98}.

In contrast, large-scale, nuclease cleavage-based structure probing experiments have recently been developed\textsuperscript{94,95}. Parallel Analysis of RNA Structure (PARS) has been used to profile mRNA secondary structures in the budding yeast, S. cerevisiae\textsuperscript{94}. Purified polyadenylated transcripts were renatured in vitro and separately treated with RNase S1 (specific for single-stranded RNA) and RNase V1 (specific for double-stranded RNA). The cleaved products from these two complementary enzymes were then analyzed using deep sequencing technology to infer single- or double-strandedness at single nucleotide resolution. Related techniques have been used to probe mRNA secondary structure in \textit{Drosophila} and \textit{C. elegans}\textsuperscript{99}. An alternative method, fragmentation sequencing (Fragseq) has been used to provide an ‘RNA accessibility profile’ on the naked RNAs from the mouse nuclear transcriptome\textsuperscript{95}. This method differs from PARS in two ways. First, Fragseq focuses on cleavage products that are 20-100 bases long, while PARS explores all the cleavage products using random fragmentation. Fragseq, thus, primarily focuses on small RNAs\textsuperscript{98}. Second, Fragseq uses only RNase P1 to cleave single-stranded RNA and reports the log ratio between the number of sequence reads obtained from the nuclease-treated sample and the untreated sample. This is done to control the occurrence of natural RNA degradation in the cell or during sample preparation. PARS, however, uses both RNase V1 and RNase S1 and reports the log ratio between the number of sequence reads obtained from the RNase V1-treated sample and the RNase S1-treated sample.

These methodologies are still undergoing development. Currently, PARS requires multiple manipulations on cellular RNA including heating and re-folding – it is unclear how the resulting product reflects the \textit{in vivo} mRNA secondary structure. Although recent reports suggest that SHAPE-like methodologies can be applied \textit{in vivo}\textsuperscript{100}, to date, genome-wide SHAPE has not been reported.
1.8 Combining experimental and computational methods

The experimentally derived RNA structure profiling data could be incorporated as a guide to computational prediction of RNA secondary structures. For example, chemical/RNase probe-based measurements of nucleotide structural conformation can be used as additional energy potentials to guide folding\textsuperscript{70,71,101}. In this case, the folding computation is biased towards RNA secondary structures that are consistent with the experimental data by assigning a large positive free-energy penalty to all possible alternatives\textsuperscript{70,71,101}. The resulting algorithm has a similar time and space complexity as default secondary structure prediction\textsuperscript{101}. It is also possible to include an additional term that reflects the inverse correlation between the SHAPE score and the base-pairing probability, and to integrate this term into the RNAstructure\textsuperscript{86} software. An alternative approach, named “sample and select”, uses experimental data to identify the correct structure among the Boltzmann ensemble of structures\textsuperscript{102}. SeqFold\textsuperscript{103} modified this approach so that only centroids of structure clusters (identified by Sfold) are considered as candidate structures. Seqfold is less sensitive to noise in experimental data than RNAstructure and “sample and select”. Binding site accessibility assessed by SeqFold has been shown to be a better predictor of \textit{in vivo} binding than RNAfold for at least some yeast RBPs; however, not all available RBPs have been assessed. However, initial results suggest that incorporating current experimental data does, indeed, improve secondary structure prediction by computational methods\textsuperscript{103}. 
1.9 RBP motif discovery algorithms that incorporate secondary structure information

Because mRNA secondary structure can either be predicted or measured, and these estimates improve the ability of computational methods to predict *in vivo* binding, it makes sense to incorporate secondary structure preferences into motif models used to scan for RBP target sites.

There are two main approaches to incorporating structural information: (1) methods that model the preferred structural context of the primary sequence motif bound by the RBP and (2) methods that explicitly model the secondary structure recognized by the RBP using stochastic context-free grammars. Table 1.2 summarizes the motif models described below.

1.9.1 Structural context-based methods

The first structural context-based method, MEMERIS, incorporates the Hackermüller-Stadler model into the popular DNA motif finding program, MEME, by annotating nucleotides according to their predicted RNA secondary structure. MEMERIS precomputes for each word (*i.e.*, *k*-mer) the probability that the word is in single-stranded context (as predicted by RNAfold), and then uses these values as priors on possible motif start positions. This adaptation changes the search so that motifs that are enriched in single-stranded regions are preferentially found. Compared to MEME, MEMERIS works better than MEME at finding RNA motifs in both artificial and in vitro datasets. As expected, MEMERIS is able to identify the correct motifs in single-stranded regions even with the existence of a stronger sequence motif embedded in a double-stranded region. MEMERIS could, in principle, easily be modified to incorporate the probability that each word is in another secondary structure context as a way of identifying, say, motifs for an RBP that binds hairpin loops; to date, such a modification has not been tested.
MEMERIS is a “generative” motif finding algorithm in that it tries to find motifs enriched among a set of bound transcripts.

Often, it is more accurate to identify “discriminative” motifs, which distinguish between sets of bound and unbound transcripts\(^47,105,106\). This approach removes the necessity to define a “background model” because it uses the unbound set. My motif-finding model #ATS and its extension in RNA-READ use this approach (described in detail in Chapters 2 and 4).

StructRED\(^104\) extends MatrixREDUCE to find RNA cis regulatory elements that are located in hairpin loops. Briefly, StructRED pre-filters all \(k\)-mers for those that are flanked by at least three bases that can pair (\(e.g., A-U, G-C, \) and \(G-U\)) and applies MatrixREDUCE to these \(k\)-mers. Unlike the two methods described above, no consideration is given to the thermodynamic stability of the stem in the naked mRNA; however, it is known that RBPs such as Vts1p can stabilize otherwise unstable loop structures\(^21-23\). StructRED correctly recovered the known binding preferences of Vts1p in yeast and its ortholog, Smaug, in flies, and discovered a number of RNA-regulatory elements in humans and flies; however, its limited representation of secondary structure elements makes it difficult to apply it to RBPs other than stem-loop binders.

RNAcontext\(^47\) is another discriminative motif model designed to identify the sequence-and structural-context preferences of RBPs from experimental binding-affinity data. RNAcontext models the relative structure preferences of an RBP to different structural contexts (\(e.g.,\) paired, hairpin loop, \(etc.)\) and this has been shown to improve the prediction performance. Its flexible alphabet for representing various structural contexts allows the method to model a wide range of RBP preferences. The input into RNAcontext consists of a set of RNA sequences, their associated structure profiles as computed by RNAplfold or Sfold, and estimates of binding affinities of the RBP of interest. Each input RNA sequence is scored using the sequence and structure context parameters. RNAcontext has been applied to \textit{in vitro} binding-affinity data for
nine RBPs\textsuperscript{16} and successfully identified both known and predicted RNA-binding preferences\textsuperscript{47}. The RBPmotif webserver implements RNAcontext\textsuperscript{107}.

1.9.2 Stochastic context-free grammar (SCFG)-based methods

There are two main methods in this category: CMfinder\textsuperscript{56} and RNApromo\textsuperscript{60}. Both fit motif models based on the covariance models (CM) that has been used to define RNA families\textsuperscript{56,60,108}. However, unlike the case for RNA families, sets of RBP target sites from different transcripts rarely have conserved sequence in paired regions, making it difficult to establish the initial alignment from which the CM iterations must be initialized. Thus, CMfinder and RNApromo use strategies based on thermodynamic stability or probabilistic consideration to establish an initial structural alignment of putative RBP binding sites from which they initialize CM. CMfinder identifies (and aligns) shared secondary structures among the minimum free energy structures of the input sequences. RNApromo replaces this initialization with non-redundant sub-structures that are overrepresented in the positive set versus the background set. One drawback to using either of these methods is that many RBPs simply bind accessible RNA – modeling the fact that the RBP binding site must be unpaired requires a particularly complex CM. These methods have been used to predict that RBPs like HuR\textsuperscript{109} and Puf3p\textsuperscript{60} (both known to bind ssRNA\textsuperscript{65,110} and to prefer accessible target sites\textsuperscript{54}) bind their targets in the stem of a hairpin loop, suggesting that they over-predict structural preferences. However, methods like these are the only ones with the capacity to model complex primary and secondary structure preferences such as those recently reported for LIN28A\textsuperscript{32} and ADAR2\textsuperscript{111,112}. That said, neither method has yet been used in this specific context; however, CMfinder has successfully identified complex structures such as riboswitches\textsuperscript{113}. 
Recently, a method called Aptamotif\textsuperscript{55} has been introduced, that adapts the iterative learning procedure of CMs to find sequence-structure motifs in SELEX-derived aptamers. As a first step, Aptamotif parses both optimal and suboptimal structures of input sequences to generate a set of loop substructures. Next, a set of seed motifs is randomly selected from the set of all loop substructures. Input sequences are scanned with these seed motifs and matching regions are aligned. The motifs with the best alignment score are retained for the next iteration. Aptamotif has been able to recover the reported binding preference of L22, a ribosomal protein that binds a long primary-sequence motif within a hairpin loop – although MEMERIS correctly identified the motif, it was not able to capture permitted gaps or the requirement for the hairpin structural context. Aptamotif’s search procedure ignores single-stranded motifs outside of loops, and the use of suboptimal structures might limit its use with long RNA sequences as the number of such structures increases exponentially with sequence length. This suggests that Aptamotif will be most useful for RBPs that have a strong requirement for specific secondary structural contexts.

1.10 Using sequence and structure conservation to find RBP binding sites

Another strategy for identification of likely RBP binding sites is to search for motifs in the 5’UTR or 3’UTRs that are surprisingly highly conserved, display a bias toward conservation when they are in the sense strand, and do not correspond to miRNA seeds. An obvious problem with this strategy is that it will fail to find binding sites for RBPs that preferentially interact with open reading frames. The approach was first used within the context of genome-wide discovery of regulatory motifs\textsuperscript{114}. Based on distinct patterns of genome-wide conservation of known motifs versus random sequences across four yeast species (\textit{S. paradoxus}, \textit{S. mikatae}, \textit{S. bayanus} and \textit{S. cerevisiae}), conservation criteria were used to discover regulatory motifs. The algorithm uses an enumeration approach to select strongly conserved motif cores and then extend or collapse these
motifs to produce candidate regulatory motifs. A similar comparative genomics analysis approach has been applied to the genomes of 12 *Drosophila* species\textsuperscript{115}. This particular method used the total branch length over which a motif is conserved, to estimate the conservation level of a motif instance. Such a scoring system is robust to comparative genomic analysis since it does not explicitly penalize missing instances, but instead rewards the motif instances in distantly related species more than ones in closely related species in order to capture neutral divergence of the motifs\textsuperscript{115}. Motifs describing primary and secondary structure preferences have been detected using a comparative method called EvoFam, which uses phylogenetic, stochastic context-free grammars\textsuperscript{116} to identify conserved, potentially regulatory, RNA structures in a 41-way genomic vertebrate alignment\textsuperscript{117}.

1.11 Overview of thesis research

The main goal of my research has been to carry out computational analyses of the target specificity and functions of RBPs in post-transcriptional regulation. RNA-protein interactions differ from DNA-protein interactions because of the central role of RNA secondary structure. RBPs bind to mRNAs either in a sequence-specific manner or by recognizing the shape of their targets. In Chapter 2 and 3, I describe my work on assessing the role of mRNA secondary structure in target recognition for both types of RBPs.

In contrast with DNA-protein interaction, high specificity to primary target sequence does not appear to be sufficient for RNA-protein interactions. Considering the dissimilarity between RNA and DNA, particularly the deeper and narrower major groove in RNA A-form helix compared to the DNA B-form helix structure, I proposed target-site accessibility as a representation of RNA structure that intuitively may capture features of the RNA structure besides that described by
linear sequence. In Chapter 2, I describe my work on assessing the role of accessibility in RBP-target interactions for more than a dozen RBPs from yeast, flies and humans.

Staufen is a fundamental member of the other class of RBPs. It is generally believed that Staufen recognizes its targets mainly via the shape of the RNA. Biochemical studies have suggested that the third dsRBD from Drosophila Staufen binds optimally to stem-loops with a 12-bp uninterrupted stem. Mutational analysis has shown that a perfect 19-bp stem embedded in a complex structure is required for human Sta1 binding to ARF1 (ADP ribosylation factor 1) mRNA. However, these perfect stems are not present in most of the Staufen targets. In Chapter 3, by analysing two RIP-Chip experiments performed in Drosophila early embryos, I describe my computational analyses of the binding preference of Staufen.

In Chapter 4, I present my work on developing and applying the RNA Regulatory Element Analysis and Discovery (RNA-READ) pipeline. RNA-READ performs both enrichment analysis of previously reported RNA motifs (RNA-READ) and de novo RNA motif discovery (RNA-READ) to identify the consensus RNA motif that best discriminates between positive and negative sets of transcripts.
Chapter 2

Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure

This chapter is derived from the following published article: X. Li, G. Quon, H.D. Lipshitz and Q. Morris. RNA (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. I did all the analysis included in this Chapter. Supplementary Files S2.1-S2.2 and the relevant codes can be found in the accompanying CD.
2.1 Abstract

While many RNA-binding proteins (RBPs) bind RNA in a sequence-specific manner, their sequence preferences alone do not distinguish known target RNAs from other potential targets that are co-expressed and contain the same sequence motifs. Recently, the mRNA targets of dozens of RNA-binding proteins have been identified, facilitating a systematic study of the features of target transcripts. Using these data, I demonstrate that calculating the predicted structural accessibility of a putative RBP binding site allows one to significantly improve the accuracy of predicting in vivo binding for the majority of sequence-specific RBPs. In my new in silico approach, accessibility is predicted based solely on the mRNA sequence without consideration of the locations of bound trans-factors; as such, my results suggest a greater than previously anticipated role for intrinsic mRNA secondary structure in determining RBP binding target preference. Target site accessibility aids in predicting target transcripts and the binding sites for RBPs with a range of RNA-binding domains and sub-cellular functions. Based on this work, I introduce a new motif-finding algorithm that identifies accessible sequence-specific RBP motifs from in vivo binding data.

2.2 Introduction

In eukaryotic cells, post-transcriptional regulation of mRNA stability, translation, localization, and splicing involves the targeting of transcripts by various RBPs that recognize cis-elements in the transcript sequence. To map out post-transcriptional networks, transcripts associated with RBPs have been identified in genome-wide assays. In many cases, these target sets are enriched for short RNA sequence motifs that reflect the sequence-binding preferences of the assayed RBPs. However, these sequence preferences do not provide sufficient specificity to distinguish the RBP-associated transcripts from unbound transcripts containing the
same short sequence motifs. While some RBPs recognize their binding sites within a hairpin loop [e.g., Vts1p\textsuperscript{22}], most mRNA-binding RBPs bind unstructured ssRNA\textsuperscript{123}, so specific RNA secondary structures are unlikely to provide the required specificity for many RBPs, thus limiting the applicability of recent algorithms developed to identify secondary structures bound by RBPs\textsuperscript{60,104}. mRNA secondary structure may instead provide specificity by sequestering potential RBP target sites within regions of double-stranded RNA, thus rendering them non-functional. This is an extension of the long held view that sequence-specific RBPs, unlike DNA-binding proteins, require at least some of their binding site to be single-stranded\textsuperscript{14} (see Chapter 1 for the detailed information).

This bias towards structurally accessible binding sites can be exploited to improve in vitro predictions of binding of RBPs to small RNAs. A role for accessibility in RBP binding has long been supported by in vitro selection\textsuperscript{124,125} and measurements of in vitro binding affinity\textsuperscript{68}. More recent, motif-finding algorithms have been developed that use measures of RNA single-strandedness to more accurately recover some RBP motifs from in vitro selection binding data\textsuperscript{53}. In all cases, computational models of RNA folding were used to predict RNA secondary structure.

However, despite its predictive value in vitro, structural accessibility has not been used to aid in the prediction of in vivo binding of RBPs to mRNA, in part because of the perceived difficulty of accurately predicting mRNA secondary structure computationally. Popular RNA secondary structure prediction methods use simplified energy models, and largely ignore the effect of co-transcriptional folding of mRNA on its secondary structure (see Kinwalker program\textsuperscript{126}). These approximations are thought to have a large impact on the accuracy of their predictions for longer RNAs such as mRNAs. However, despite these deficiencies, structural accessibility calculated by these programs does predict the in vivo binding sites of miRNAs\textsuperscript{63,64,127} and siRNAs\textsuperscript{61}, thus demonstrating that these methods do predict single-strandedness with some degree of accuracy.
Nonetheless, using target-site accessibility to predict in vivo RBP binding has remained largely untested because of obvious differences between RNA binding by RBPs versus miRNAs/siRNAs. Unlike these ncRNAs, many RBPs function in the nucleus where mRNA secondary structure may be much more constrained by large heterogeneous ribonucleoprotein complexes (hnRNPs) associated with the transcript that are displaced during the export of the mRNA to the cytoplasm or during the first round of translation. It has also been suggested, based on the presence of RNA helicases and potential RNA chaperones within hnRNPs, that mRNA secondary structure undergoes extensive remodeling to facilitate RBP binding. Furthermore, although miRNAs and siRNAs compete for the same binding interface as mRNA secondary structure, RBPs can bind RNA through a variety of interfaces, some of which may only require small disruptions in A-form helical structure to expand the major groove and thereby permitting recognition of bases flanking the disrupted region. Under this circumstance, only a subset of the bases within the sequence-specific binding site need be unpaired.

The recent availability of mRNA target sets for a large number of RBPs from yeast, flies and humans has allowed us to assess the impact of structural accessibility on sequence-specific binding of a diverse set of RBPs that carry a variety of RNA-binding domains and participate in a number of subcellular functions. Through a systematic analysis, I demonstrate that target site accessibility, predicted based on intrinsic mRNA secondary structure, plays a general role in RBP binding. Incorporating target site accessibility into computational models of sequence-specific RBP binding yields a statistically significant overall improvement in their ability to predict the outcome of large-scale assays of RBP-mRNA interactions for the majority of sequence-specific RBPs.
2.3 Results

To investigate the role of mRNA secondary structure on RBP binding, I compiled data on the \textit{in vivo} mRNA targets of a set of 30 eukaryotic RBPs from \textit{Saccharomyces cerevisiae}, \textit{Drosophila melanogaster} and humans derived from RNP immunoprecipitation microarray (RIP-chip) co-purification assays. Details of each dataset are available in Methods and Supplementary File S2.1. I assessed the impact of mRNA secondary structure on putative RBP binding sites by scoring their accessibility and determining whether more accessible target sites were more likely to be bound. I define “target site accessibility” as the probability that the entire target site is unpaired as estimated by a computational method, RNAplfold\textsuperscript{76}, which considers the relative stabilities of all possible secondary structures containing the site and its flanking sequence (see Methods for details).

2.3.1 Target site accessibility predicts mRNA targets of Pumilio and Puf3p

I began the investigation by examining RIP-chip-derived target sets for \textit{Drosophila} Pumilio\textsuperscript{50}, a well-studied protein with conserved RNA binding specificity, and \textit{S. cerevisiae} Puf3p\textsuperscript{49}, the likely yeast ortholog of Pumilio. These two proteins, along with the Fem-3-binding factor in \textit{Caenorhabditis elegans}, share a conserved RNA-binding domain consisting of eight repeats of the Pum-homology domain (PumHD). To date, all known Puf proteins bind their targets through this domain and subsequently regulate the stability and/or translation of these targets\textsuperscript{119,129-131}. Structural\textsuperscript{20,110,132}, small-scale\textsuperscript{133,134}, and large-scale\textsuperscript{49,50} studies are consistent with a conserved single-stranded consensus binding sequence, UGUHAUA, for Pumilio and Puf3p (H indicates that A, C, or U is permitted).
To evaluate the role of accessibility in Puf3p and Pumilio targeting in vivo, I defined a set of mRNAs likely to be bound by fly Pumilio and yeast Puf3p based on their relative enrichment in the bound fraction of mRNA using FDR cutoffs established in the original studies. As a negative control, I also defined a set of mRNAs expressed under the queried conditions, and thus available for binding, but that were not enriched in the co-immunoprecipitated fraction (see Supplementary File S2.1 for details). I called these transcripts “unbound”. All experimentally validated target sites to date for Pumilio and Puf3p occur in the 3’UTR, so I scanned the 3’ UTRs of mRNAs in the bound and unbound sets and identified those that contained a match to the UGUAHAUA consensus. As expected, a larger proportion of bound transcripts contained a 3’UTR match to UGUAHAUA (Puf3p 75% versus 9%, $P < 3.0 \times 10^{-94}$, Pumilio 51% versus 12%, $P < 1.0 \times 10^{-81}$, Fisher’s Exact Test). However, there were more unbound transcripts with matches than bound ones (158 vs. 246 in yeast, 241 vs. 482 in fly) (Figure 2.1A,B). As such, target-site recognition for Pumilio and Puf3p cannot be explained by RNA sequence preference alone. To determine whether target site accessibility could distinguish unbound transcripts from bound ones, I estimated the probability that each match to UGUAHAUA was single-stranded, using a computational method that predicts target site accessibility based only on RNA sequence flanking the target site (see Methods), and compared the accessibility of matches in the 3’UTRs of bound and unbound transcripts. Multiple consensus sites within the same 3’UTR may increase the affinity of the RBP for the mRNA; to control for this, I only compared bound and unbound transcripts with the same number of matches. Figure 2.1C and D contain the results for transcripts with a single match in their 3’ UTR, which constitute the vast majority of transcripts. The results for transcripts with multiple matches are similar (Table A3.1).

As Figure 2.1 shows, the median accessibility of sites in bound mRNAs was almost two-fold higher than in unbound mRNA in both yeast (Figure 2.1C, inset) and fly (Figure 2.1D, inset). Furthermore, ROC analysis demonstrated that target-site accessibility is a statistically significant predictor of co-immunoprecipitation of a transcript with Pumilio or Puf3p (Figure 2.1C and D;
Puf3p AUROC = 0.74, $P = 3 \times 10^{-14}$; Pumilio AUROC = 0.65, $P = 3 \times 10^{-9}$, Wilcoxon-Mann-Whitney test). These results demonstrate that target site accessibility plays a role in RBP binding, and thus target mRNA selection, by Pum and Puf3p.

We also performed a similar test for one of the human homologs of Pumilio, human Pum1, using RIP-chip-derived target sets from Morris et al.\textsuperscript{135} (Figure 2.2). Although in this case, there are more bound than unbound transcripts with 3’ UTR copies of UGUHAUA (Figure 2.2A), accessibility remained a statistically significant predictor in determining Pum1 binding upon comparison of bound versus unbound transcripts containing the same number of consensus sites in their 3’UTRs (Figure 2.2B and Table 2.1).
Figure 2.1. Puf3p and Pumilio consensus binding sites have higher accessibility in 3’UTRs of their bound mRNA targets.

(A) and (B) While the consensus matches were significantly enriched in the set of bound transcripts for yeast Puf3p and fly Pumilio, more unbound transcripts contained consensus matches than bound ones (158 vs 246 for yeast Puf3p (A), 241 vs 482 for fly Pumilio (B)). (C) and (D) Comparison of site accessibility of transcripts co-immunoprecipitating (co-IPing) with Puf3p (C) and Pumilio (D) and those co-expressed but not co-IPing. All compared transcripts have only a single copy of the Puf3p / Pumilio consensus UGUAHAUA (H matches A, C or U) in their 3’ UTRs (132 bound and 235 unbound transcripts for Puf3p; 201 bound and 414 unbound transcripts for Pumilio). The ROC curve (solid line) plots the sensitivity (i.e., the proportion of bound transcripts recovered; vertical axis) against [1 – specificity] (i.e., the proportion of unbound transcripts recovered; horizontal axis) as the accessibility threshold is adjusted from the highest to the lowest. Inset: median site accessibility for the bound set (dark grey bar) and the unbound set (light grey bar). Error bars represent the 95% confidence interval of the median calculated using 5,000 bootstrap samples. P-values were calculated using the Wilcoxon-Mann-Whitney Rank Sum test.
Figure 2.2. Pum1 sequence consensus site have higher accessibility in 3’UTRs of their mRNA targets.

(A) While the consensus matches were significantly enriched in the set of 301 bound transcripts for human Pum1, there were still 191 unbound transcripts that contained consensus matches. (B) Comparison of site accessibility of transcripts co-immunoprecipitating (co-IPing) with Pum1 and those co-expressed but not co-IPing. All compared transcripts have only a single copy of the Pum1 consensus UGUHAUA (H matches A, C or U) in their 3’UTR (185 bound and 158 unbound transcripts). The ROC curve (blue solid line) plots the sensitivity (i.e., the proportion of bound transcripts recovered; vertical axis) against [1 – specificity] (i.e., the proportion of unbound transcripts recovered; horizontal axis) as the accessibility threshold is adjusted from the highest to the lowest. Inset: median site accessibility for the bound set (green bar) and the unbound set (red bar). Error bars represent the 95% confidence interval of the median calculated using 5,000 bootstrap samples. P-values were calculated using the Wilcoxon-Mann-Whitney rank sum test.
<table>
<thead>
<tr>
<th></th>
<th>AUROC</th>
<th>P-value</th>
<th>Bound hits</th>
<th>Unbound hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Puf3p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 site in 3'UTR</td>
<td>0.74</td>
<td>3.0 x 10^{-14}</td>
<td>132</td>
<td>235</td>
</tr>
<tr>
<td>2 sites in 3'UTR</td>
<td>0.80</td>
<td>0.006</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>Pumilio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 site in 3'UTR</td>
<td>0.65</td>
<td>2.6 x 10^{-9}</td>
<td>201</td>
<td>414</td>
</tr>
<tr>
<td>2 sites in 3'UTR</td>
<td>0.62</td>
<td>0.06</td>
<td>34</td>
<td>59</td>
</tr>
<tr>
<td>Pum1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 site in 3'UTR</td>
<td>0.61</td>
<td>4.7 x 10^{-4}</td>
<td>185</td>
<td>158</td>
</tr>
<tr>
<td>2 sites in 3'UTR</td>
<td>0.62</td>
<td>0.07</td>
<td>87</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 2.1: Target site accessibility predicts mRNA targets of Pumilio and Puf3p.

(A) Comparison of site accessibility for Puf3p sites, stratifying corresponding bound and unbound sets into the groups with same number of UGUAAUA sites in 3'UTR. Note that I only show the comparisons with statistical power (P-value < 0.05). When there was more than one hit in the 3'UTR of a transcript, accessibility for this transcript was presented by sum of the accessibility for all the sites. (B) and (C) Per A. for fly Pumilio (B) and human Pum1 (C). AUC score is the area under ROC curve. The P-value was calculated by using Wilcoxon-Mann-Whitney rank sum test.
2.3.2 Accessibility improves motif-based mRNA target prediction for diverse RBPs

Having established proof-of-principle that target site accessibility, predicted on the basis of mRNA sequence alone, has a measureable impact on RBP binding in vivo, I sought to determine the generality of this observation by assessing the impact of target site accessibility for RBPs with a diverse range of RNA-binding domains, sequence-binding preferences, and subcellular functions. To do so, I compiled RIP-chip data and consensus sequence motifs for additional RBPs from yeast and human. Of 18 such RBPs (including Puf3p, Pumilio and Pum1), I removed three (Pab1p, Nsr1p, Nrd1p) whose bound sets were not significantly enriched for the reported consensus sequence (Wilcoxon-Mann-Whitney P-value > 0.05) and one RBP (Ssd1p) for which only seven transcripts matched the consensus. Thus, I were left with 14 likely sequence-specific RBPs with a large enough number of putative target mRNAs for my analysis. Although the fact that these RBPs are sequence-specific suggests that at least some portion of their binding site needs to be structurally accessible, many of these RBPs lack crystal structures, so it is not clear how much of the site needs be accessible or whether the structural accessibility of the site can be predicted by computational folding of the mRNA sequence without consideration of the influence of trans-factors on the mRNA’s secondary structure. For each of these RBPs, I again used the relative enrichment among mRNAs co-purifying with the RBP, as measured using the RIP-chip assay, to define its bound mRNA transcripts and a set of unbound mRNA transcripts that were co-expressed with the RBP but showed no evidence of being bound (see Supplementary File S2.1 for details).

Some of the RBP consensus sites matched in a large number of positions, making it difficult to directly compare transcripts with the same number of potential target sites. I therefore adopted a new and more general analytical procedure (described schematically in Figure 2.3) that
compared all bound transcripts to all unbound ones and scanned the entire mature mRNA sequence for binding sites. To assess the predictive value of target site accessibility, I assigned each transcript a score equal to the sum of the accessibilities of sites in the transcript, and then evaluated how well that score distinguished bound and unbound transcripts. I call this score the "total accessibility" and abbreviate it by "#ATS" because it is equal to the expected number (#) of Accessible Target Sites per transcript. To control for the fact that transcripts with more RBP binding sites were more likely to be bound, I also assessed how well the number of target sites (which I abbreviate by “#TS”) in a transcript predicted whether or not it was bound. As before, I evaluated the predictive accuracy of both #ATS and #TS using AUROC. I tested for a significant difference between the two AUROCs by combining a permutation test with the Delong-Delong-Clarke-Pearson procedure¹³⁶ (see Methods for more details).
The flowchart displays the procedure for evaluating accuracy at distinguishing bound and unbound sets of mRNA using either #ATS- or #TS-based scoring of an RBP consensus sequence. For each RIP-chip data set, transcripts were sorted in decreasing order by their relative enrichment among mRNAs co-purifying with the RBP. I defined those with relative enrichment larger than the ‘positive threshold’ to be the “bound” set of transcripts and those with relative enrichment smaller than the ‘negative threshold’ to be the “unbound” set of transcripts. In this way, it was guaranteed that the transcripts in the unbound set were co-expressed with the RBP. I then identified all consensus-sequence matches (which I call “target sites”) in each transcript and removed transcripts with no target sites. I then ranked transcripts in decreasing order of number of target sites and used this ranking to calculate the #TS AUROC. To calculate #ATS, I first calculated the accessibility of each target site. This calculation considered all possible secondary structures, weighted according to their stability, so even sites that were single-stranded or paired in the most probable secondary structure (as displayed) could have a value less than 1 or larger than 0, respectively. I ranked transcripts in decreasing order by the sum of the accessibilities of their target sites (i.e., #ATS) and calculated the associated AUROC.
Considering only the transcripts that contained at least one copy of the consensus sequence, for 13 of the 14 RBPs there was a statistically significant increase in #TS or #ATS among the bound transcripts (P < 0.05, Wilcoxon-Mann-Whitney) compared to unbound. These 13 RBPs, along with their AUROCs, are displayed in Figure 2.4. For 10 of 13 RBPs, there was a statistically significant increase in AUROC when #ATS versus #TS was used to predict whether a transcript would be bound. In some cases the increase in AUROC was quite large while in other cases it was more modest. However, is should be noted that because random performance for AUROC is 50%, absolute increases in AUROC translate into relative decreases in error that are at least twice as large (where error is measured as 100% - AUROC). For example, although the average absolute increase in AUROC for all motifs in Figure 2.3 is 9.3%, the average relative decrease in error is 22.1%. Notably, in no case did #ATS have an AUROC significantly smaller than the AUROC for #TS.

Target site accessibility predicted not only the targets of proteins thought to bind unstructured ssRNA (e.g., the Puf family of proteins) but also the targets of Vts1p, a stem-loop-binding protein that recognizes loops containing CNGG(N)_{0-3}. Indeed, my procedure was sensitive enough to detect the fact that Vts1p binds loops of length four bases or more: #ATS had a significant improvement in accuracy over #TS when used with the CNGG motif and not a CNGGN motif because the latter would score four-base loops bound by Vts1p as inaccessible.

In summary, target site accessibility is a statistically significant predictor of in vivo binding for 71% (10 of 14) of the sequence-specific RBPs tested.
Figure 2.4. Target site accessibility predicts *in vivo* binding for a diverse range of RBPs.

Bar graphs compare the accuracy of #ATS and #TS at predicting bound transcripts based on a given consensus. To the left of the bar graph each row is labeled by the RBP, the associated consensus sequence used for classification, and a cartoon indicating the species of origin (yeast, fly, or human). Some RBPs have multiple reported consensus sequences; these are grouped and indicated by a vertical bar. To the right of the bar graph, for each RBP, I show its known sub-cellular localization and its known RNA-binding domains (using SMART domains). The left localization column indicates nuclear localization (if any) as indicated by Hn (hnRNP) or Nu (nucleus); the right localization column describes cytoplasmic localization as indicated by Cy (cytoplasm), ER (endoplasmic reticulum), Mi (mitochondrion), Ri (ribosome), SG (stress granule). Supplementary File S2.1 contains the evidence for the reported localization and domains. The statistical significance of differences between #ATS AUROC (green bars) and #TS AUROC (yellow bars) was calculated using the Delong-Delong-Clarke-Pearson procedure and is indicated by asterisks: * indicates $P < 0.05$, ** indicates $P < 0.01$, *** indicates $P < 10^{-4}$. Exact $P$-values are in Supplementary File S2.2.
2.3.3 Improvement due to accessibility is not explained by nucleotide composition biases

I performed a number of computational controls to confirm that the observed role of target site accessibility was not due to other potential properties of functional binding sites. One possible alternative explanation for the observed differences is that functional RBP binding sites are in regions of biased, low-order nucleotide composition. For example, many cis-regulatory mRNA elements are located in 3’ UTRs, which tend to be AU-rich. This AU-richness of flanking sequence has been suggested as an explanation for the predictive value of accessibility for miRNA binding\(^{137}\). However, if this were the case for RBPs, then one would expect either (a) that the effect would disappear if my analysis was restricted by only scanning sites in the 3’UTR or (b) that the flanking regions around target sites in bound transcripts would be biased towards a single type of dinucleotide. However, target site accessibility remained a strong predictor of in vivo binding when I restricted my scans to 3’UTRs (Figure 2.5). Furthermore, although there were statistically significant differences in the dinucleotide composition of sequence flanking sites in bound transcripts, the enriched and depleted dinucleotides depended on the consensus sequence and tended to favor dinucleotides less likely to pair with the consensus (Figure 2.6). Indeed, most dinucleotides were significantly enriched for some RBPs and significantly depleted for others. I also confirmed that the increase in predictive power was not due to the sequence composition of the consensus; when I repeated my analysis using the reverse consensus (e.g. CNGG became GGNC), I only saw a significant increase for #ATS when the reversed consensus was a strong match to the forward consensus (Figure 2.7).
Figure 2.5. 3'UTR target site accessibility predicts in vivo binding.

Results are presented as in Figure 2.3, but only target sites within the 3'UTRs of transcripts were used to calculate #TS and #ATS. Plus signs on the bars indicate that the AUROC is significantly different from random (Wilcoxon-Mann-Whitney, $P < 0.05$). Exact $P$-values are in Supplementary File S2.2 in the CD.
Figure 2.6. Differences in dinucleotide composition around putative RBP binding sites between bound and unbound transcripts.

(A) Heat map showing the t-statistic of the difference in di- and single-nucleotide frequencies in 40 bases upstream and downstream of the target site. Rows indicate the RBP and its consensus binding sequence motifs (in IUPAC representation) used to identify target sites. Columns indicate single versus di-nucleotide. Rows and columns were ordered based on 2-dimensional hierarchical clustering. Those t-statistics with absolute value less than two are not statistically significant at $\alpha = 0.05$ and are set to zero; those with an absolute value greater than four remain statistically significant after a Bonferroni correction and are thresholded at 4 or -4, as appropriate. (B) As for a. but using 20 bases up- and downstream.
Figure 2.7. Comparative predictive accuracy of #TS and #ATS when reverse consensus motifs are scored.

As per Figure 2.4, but consensus sequence motifs were reversed before scoring (as displayed). Plus signs on bars indicate the AUROC was significantly different from random (Wilcoxon-Mann-Whitney, \( P < 0.05 \)). Exact \( P \)-values are in Supplementary File S2.2 in the CD.
2.3.4 Many RBPs require the entire target site to be accessible

I next sought to determine whether different approaches to calculating target site accessibility had an impact on the performance of my assay. Different approaches make different assumptions about the role of target site accessibility in RBP binding. The method I employed, as proposed by Hackermüller and co-workers, requires the whole site to be unpaired for the protein to bind. However, other methods, including the EF option of MEMERIS, and some methods used to predict target site accessibility for miRNAs, allow target sites to be partially paired. To determine which of these approaches most accurately predicts in vivo RBP binding, I compared transcripts scored by #ATS based on the accessibility of the whole target site and those scored by #ATS when target site accessibility was approximated by either the average or the minimum single base accessibility of all bases in the target site.

The estimates of target site accessibility from these two methods (single base and whole site) diverge most when the target site was partially paired. For example, if in all stable mRNA secondary structures exactly half of the bases in the target site were paired then the average single base accessibility for that site would be 0.5 but the accessibility of the whole target site would be 0. On the other hand, if the target site was completely unpaired in some structures and completely paired in others then both the average single base accessibility and the whole site accessibility would be exactly the same (Figure 2.8A).

When I compared the predictive accuracy of #ATS calculated using whole-site, average-single-base, and minimum-single-base accessibilities (Table 2.2), I found that for six proteins (Puf3p, Puf4p, PTB, HuR, Khd1p, Vts1p), both of the single-base approximations significantly decreased predictive accuracy (Figure 2.8B). Four of these six [Puf3p, Puf4p, PTB, and Vts1p] have solved co-crystal structures showing that the RBP binds to a completely unpaired
target site, and there is other evidence to suggest that the entire HuR binding site must be unpaired\textsuperscript{65,124,125}. My data thus suggests that Khd1p will also require its entire binding site to be unpaired. It should be noted that I found no instances where either of the single-base approximations significantly improved accuracy compared with the whole-site method.
Figure 2.8. Target site accessibility is a better predictor than average/minimal accessibility of single bases in the target site.

(A) Diagrams represent examples of how secondary structure leads to differences in the calculated target site accessibility when the calculation is for the entire site (green), minimal single-base accessibility (magenta), or average single-base accessibility (light blue). In each case, two equally stable structures are shown and the numbers represent accessibility calculated for a four-base site assuming each secondary structure is equally probable. (B) As per Figure 2.3 except that light blue and magenta bars show AUROCs for #ATS scoring when the target site accessibility is replaced with the average and minimal accessibility of all single bases in the target site. Exact P-values are in Supplementary File S2.2 in the CD.
<table>
<thead>
<tr>
<th>RBP</th>
<th>Consensus seq</th>
<th>Average sb % change of AUROC</th>
<th>Min sb % change of AUROC</th>
<th>Max access % change of AUROC</th>
<th>Flank % change of AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Msl5</td>
<td>UACUAAC</td>
<td>-1.74</td>
<td>1.47</td>
<td>1.22 *</td>
<td>10.15</td>
</tr>
<tr>
<td>PuF4</td>
<td>UGUAUAUAUA</td>
<td>-2.85</td>
<td>-1.49</td>
<td>-0.16</td>
<td>0</td>
</tr>
<tr>
<td>PuF4</td>
<td>UGUAHMNUA</td>
<td>-10.02 ***</td>
<td>-5.98 **</td>
<td>-1.00 *</td>
<td>0</td>
</tr>
<tr>
<td>PuF3</td>
<td>CNUGUAHAUA</td>
<td>-10.02 *</td>
<td>-9.84 *</td>
<td>-0.39</td>
<td>0.55</td>
</tr>
<tr>
<td>PuF3</td>
<td>UGUAHAUA</td>
<td>-6.52</td>
<td>-6.64</td>
<td>-1.77 **</td>
<td>3.47</td>
</tr>
<tr>
<td>Pumilio</td>
<td>UGUAHAUA</td>
<td>-4.93 *</td>
<td>-2.34</td>
<td>-0.63</td>
<td>0</td>
</tr>
<tr>
<td>PTB</td>
<td>UCUUC</td>
<td>-4.16 ***</td>
<td>-2.31 ***</td>
<td>-2.48 *</td>
<td>0</td>
</tr>
<tr>
<td>PTB</td>
<td>UCUU</td>
<td>-3.50 ***</td>
<td>-2.01 ***</td>
<td>-2.24</td>
<td>0</td>
</tr>
<tr>
<td>PTB</td>
<td>YCU</td>
<td>-3.56 ***</td>
<td>-1.79 ***</td>
<td>-0.56</td>
<td>0</td>
</tr>
<tr>
<td>HuR</td>
<td>NNUUNNUUU</td>
<td>-4.45 ***</td>
<td>-2.62 ***</td>
<td>-4.58 **</td>
<td>0</td>
</tr>
<tr>
<td>Khd1</td>
<td>CNNCNN</td>
<td>-5.75 ***</td>
<td>-3.31 ***</td>
<td>-5.73 ***</td>
<td>0</td>
</tr>
<tr>
<td>Nab2</td>
<td>DRARMGMGD</td>
<td>-3.39</td>
<td>-1.44</td>
<td>-3.87 *</td>
<td>0.17</td>
</tr>
<tr>
<td>Yll032c</td>
<td>AAUACCY</td>
<td>2.50</td>
<td>-3.41</td>
<td>-0.29</td>
<td>5.49</td>
</tr>
<tr>
<td>Vts1</td>
<td>CNGG</td>
<td>-4.87 *</td>
<td>-3.13 *</td>
<td>4.72</td>
<td>0</td>
</tr>
<tr>
<td>Vts1</td>
<td>CNGGN</td>
<td>-0.28</td>
<td>-1.55</td>
<td>-0.43</td>
<td>0</td>
</tr>
<tr>
<td>Pub1</td>
<td>HUUUUUHHWY</td>
<td>-1.01</td>
<td>5.46</td>
<td>-1.78</td>
<td>0.66</td>
</tr>
<tr>
<td>Pum1</td>
<td>UGUAHAUA</td>
<td>-3.12</td>
<td>1.43</td>
<td>-3.19 ***</td>
<td>6.37</td>
</tr>
<tr>
<td>PuF2</td>
<td>UAAUAAYW</td>
<td>-0.63</td>
<td>1.60</td>
<td>-5.82 **</td>
<td>8.50</td>
</tr>
</tbody>
</table>

Table 2.2. The impact of other variations in the calculation of target site accessibility.

Each element corresponds to the percent change in AUROC compared to the AUROC for #ATS when transcripts are scored with the new method. For example, if the new AUROC were x the percent change would be 100% * (x - #ATS AUROC) / (#ATS AUROC). The methods are abbreviated as follows: “average sb” is when the site accessibility in #ATS is replaced with the average single-base accessibility in the site; "min sb" replaces site accessibility with minimum single-base; "max access” replaces the sum of the accessibilities of all sites in the transcript with the maximum; "flank" adds flanking region accessibility to the site accessibility. The statistical significance of differences between the #ATS and each other AUROC was calculated using our combined permutation test and Delong-Delong-Clarke-Pearson procedure, with FDR correction, and is indicated by asterisks: * FDR < 0.1, ** FDR < 0.01, *** FDR < 0.001. Exact P-values are in Supplementary File S2.2 in the CD.
2.3.5 The impact of other variations in the calculation of target site accessibility

I evaluated two other methods for calculation of target site accessibility. First, when transcripts were scored using the maximum accessibility of the target sites in the transcripts, AUROC scores almost always decreased relative to #ATS, often significantly (Table 2.2; the only exceptions were Vts1p and Msl5p). These results suggest that multiple accessible sites in the same transcript contribute to binding. Second, I investigated whether also considering the accessibility of sequence flanking the target site helped predict binding. Unlike the case for miRNAs, I found that for RBPs there was little improvement in AUROC when a measure of flanking region accessibility was added to target site accessibility (Table 2.2).

2.3.6 In vivo sequence motif finding using accessibility

I next attempted to recover the RBP sequence-binding preferences from the in vivo data using motif-finding methods that either did or did not incorporate target site accessibility. I carried out these analyses for two reasons: first, I wanted to ensure that the increased predictive accuracy of accessibility did not arise from how the RBP consensus sequences were originally defined; second, I wished to assess whether incorporation of accessibility as a feature improved either the accuracy or the statistical power of RBP motif finding based on in vivo co-purification data. I performed this motif-finding analysis on all yeast RBPs in my collection: these included 14 with previously associated motifs and 12 with no associated sequence motifs (Bfr1p, Cbc2p, Cbf5p, Gbp2p, MRN1p (also known as Ypl184c), Nab3p, Nab6p, Nop56p, Npl3p, Puf1p, Scp160p, and Tdh3p).
I used a discriminative motif-finding procedure that attempted to identify consensus sequences that, when scored with either #ATS or #TS, best distinguished bound and unbound transcripts by being more often present and/or having either more sites (for the #TS-derived motif) or more accessible sites (for the #ATS-derived motif) among the bound transcripts (see Methods for details). In Figure 2.9, I report two motifs for each RBP: one based on #TS and the other #ATS. These motifs were derived from the entire set of bound and unbound transcripts. However, the AUROCs that I report for #ATS and #TS were calculated on held-out data using a cross-validation training procedure that I employed to avoid over-fitting (see Methods). Motifs learned during cross-validation were similar to those reported in Figure 2.9 (Supplementary File S2.1 in the CD contains all learned motifs; note I do not display consensus sequence motifs for the nine RBPs for which my procedure was unable to find motifs whose predictive accuracy was significantly better than random).

Both #ATS-based and #TS-based motifs were consistent with previously determined binding preferences for all nine yeast RBPs displayed in Figure 2.4. Also consistent with Figure 2.4, the AUROC of the #ATS motif on held-out data was significantly higher than that of the #TS for seven of the nine RBPs reported therein. As expected, the #ATS-trained motif for Vts1p recovered the bound loop sequence and achieved a higher AUROC than the #TS-trained motif which attempted to model the stem sequence. For eight of the nine RBPs, there was also a significant increase in AUROC when I used #ATS to score the held-out data based on the #TS-trained motif (the only exception was Vts1p). Thus, even motifs trained to maximize the #TS-based score remained more predictive of RBP binding when scored with #ATS demonstrating that target site accessibility increased accuracy regardless of how the sequence motif was derived.

Figure 2.9 also contains two RBPs not previously associated with motifs, Gbp2 and MRN1/Ypl184c. For these RBPs, the #ATS-derived and #TS-derived motifs were similar and
there was a statistically significant increase in accuracy when scoring either of these motifs using #ATS thus following the same pattern as most of the RBPs already associated with motifs. These results suggest that these two RBPs also bind unstructured ssRNA. Figure 2.10 displays motifs for the six other RBPs: for three of these RBPs (Bfr1, Scp160 and Tdh3), the #ATS-derived motif was contained within the motif discovered by #TS and the AUROC for #ATS scoring based on the #TS-derived motif was smaller than the two other AUROCs. This pattern mirrors Vts1p, suggesting that the additional bases in the #TS-derived motif are likely to be inaccessible.

In summary, my motif analysis demonstrated two things. First, the ability of target site accessibility to improve accuracy (as shown in Figure 2.4) is not an artifact of how the previously reported RBP sequence motifs were defined; even motifs trained to maximize #TS AUROC in my assay underwent a significant improvement in accuracy when scored using #ATS. Second, the #TS- and #ATS-derived motifs were very similar for RBPs that bind unstructured ssRNA and for these RBPs, I often observed no significant difference in AUROCs when the two motifs were scored using #ATS. This observation suggests that mRNA secondary structure functions primarily to sequester non-functional matches to an RBP’s sequence preferences rather than to reveal binding sites for RBPs with highly degenerate sequence preferences.
Figure 2.9. RBP motifs optimized to distinguish bound versus unbound transcripts.

Each RBP is shown associated with two motifs: the #ATS-derived motif with the highest AUROC (in green box) and the #TS-derived motif with the highest AUROC (in yellow box) after motif finding was performed on the complete set of bound and unbound transcripts. Grey background indicates overlap of manually aligned regions of the #ATS and #TS motifs for the same RBP. The bar graphs display median AUROC over 30 held-out test sets for motifs trained to maximize #ATS AUROC and scored with #ATS (green bar), trained to maximize #TS AUROC and scored with #TS (yellow bar), and trained to maximize #TS AUROC but scored with #ATS (orange bar). I assessed P-values for differences in distributions of 30 AUROCs on matched test sets using the Wilcoxon sign-rank test. Italicized RBP names indicate the RBP has a previously defined consensus sequence, and bold italics indicate a significant increase in #ATS reported in Figure 2.4. The P-value threshold is indicated as for Figure 2.4 and exact P-values are available in Supplementary File S2.2 in the CD. Sub-cellular location and RBD domains are displayed for RBPs not represented in Figure 2.4.
Figure 2.10. Learned in vivo motifs for six yeast RBPs.

As per Figure 2.9 except this Figure contains the motifs for the yeast RBPs not shown in Figure 2.4.
2.4 Discussion

I have demonstrated that binding-site accessibility has a significant impact on mRNA target selection for 12 of 14 RBPs (86%) with previously determined sequence-binding preferences. Also, using a novel discriminative in vivo motif finding approach that incorporates target site accessibility, I were able to identify two additional RBPs, Gbp2 and MRN1, that are likely to bind unstructured ssRNA. Together, these 14 RBPs include five different classes of RNA-binding domains (RRM, KH, Pum-repeats, SAM, C2H2-Zn-finger), are not biased towards either nuclear or cytoplasmic function, and include examples of RBPs known to bind both unstructured ssRNA and loop sequences. Thus accessibility predicts target selection by RBPs with a diverse set of RNA-binding domains that bind within different secondary structure contexts and that have different subcellular locations of binding.

I have also identified some differences between features of RBP binding sites and miRNA and siRNA binding sites. First, as previously reported for in vitro binding, allowing partial pairing of putative RBP binding sites significantly reduces in vivo predictive accuracy for six RBPs. Five of these RBPs have previously been reported to require their entire binding site to be unpaired, strongly suggesting that the sixth, Khd1, will have a similar requirement. It should be noted that I never observed an advantage to allowing partial pairing. Also, unlike miRNA binding, requiring flanking sequence also to be accessible never significantly improved predictive accuracy.

My observations, taken together with similar observations on the role of mRNA secondary structure in small regulatory RNA targeting, demonstrate that accessibility plays a role in target selection throughout the lifetime of an mRNA. Because in all cases, structural accessibility was predicted using methods that only consider the mRNA sequence, these data suggest that
intrinsic mRNA secondary structure forms prior to trans-factor binding and constrains subsequent binding events at all levels of post-transcriptional regulation. These data also provide a possible mechanism by which the clustering of target sites in the transcript increases the likelihood of RBP binding\textsuperscript{138-140}: many of the RBP consensus motifs that I considered do not form stable RNA secondary structures when concatenated. Thus, site clustering, in addition to providing more target sites for the RBP, cooperatively enhances the accessibility of sites in the cluster. Preferential RBP binding at accessible target sites also provides a mechanism that explains why accessibility modulates the \textit{cis}-regulatory impact of known splicing enhancer and suppressor elements on nearby splice sites\textsuperscript{53}.

My data predict that HuR, Puf3p, and Puf4p require their whole binding site to be accessible and are thus consistent with previous \textit{in vitro} studies for HuR\textsuperscript{16,65,124,125} and solved co-crystal structures for Puf3p\textsuperscript{110} and Puf4p\textsuperscript{132}. However, my observations appear to conflict with recent reports for HuR\textsuperscript{109} and Puf3p and Puf4p\textsuperscript{60} in which the binding sites I propose for these RBPs are predicted to be partially paired within a hairpin. There are a number of possible explanations for this disparity. First, it is possible that these proposed hairpins do not form because they are not energetically favored. The algorithms used to identify these stem-loops, COVE\textsuperscript{58} and RNApromo\textsuperscript{60}, employ Covariance Models (CM) to predict RNA secondary structure. CMs only consider existence of possible pairings and consider neither their thermodynamic stability nor the impact of flanking sequence on the predicted structure. Indeed, as stated in the user guide to COVE, “covariance models routinely overpredict [RNA secondary] structure, because a) they don't look for Watson-Crick complementarity and b) it is often statistically advantageous for the model to pair as many positions as possible” [Guide.tex\textsuperscript{141} accessed from ftp://selab.janelia.org/pub/software/cove/cove-2.4.4.tar.Z on February 7, 2010]. Another explanation may that the hairpins do form but are comparatively less stable in bound than unbound mRNA. If true, this would be consistent with my observations because my model only considers the relative, not the absolute, accessibility of sites in bound mRNAs.
My results have important consequences for large-scale analyses of RNA-protein interactions and RNA processing, both experimentally and computationally. I have shown that target-site accessibility almost always increases – and never significantly decreases – the ability to predict sequence-specific RBP binding to mRNAs. I have also found that rewarding a transcript for containing multiple target sites improves predictive accuracy. Thus, methods that attempt to identify RBP target sites in mRNAs or to infer regulatory networks should be augmented with target site accessibility data for all potential binding sites.

My models do not perfectly reproduce the in vivo binding data, suggesting that there remains room for further improvement. Although some of the errors that my models make may be due to external factors, such as noise in the original experimental assay or my inability to precisely recover the mRNA transcript sequence targeted by the RBP, many errors are likely to be due to in vivo regulatory mechanisms that are not captured by my model. First, my model considers the accessibility of target sites in an mRNA but does not take into account whether or not that site is associated with a particular element of RNA secondary structure like a hairpin loop. I addressed this point in Chapter 4. Second, I did not model possible competition for binding sites by other trans-factors or cooperative binding with different trans-factors. These factors could, for example, be responsible for my poor performance at predicting Pum1 sites because Pum1 has similar binding preferences to Pum2 and there is an enrichment of microRNA binding sites around Human Pum sites. Third, the method that I used to predict target site accessibility considers neither the effects on RNA secondary structure of other bound trans-factors nor long-range interactions within the mRNA (e.g., 143).

Despite these caveats, my analysis demonstrates that mRNA secondary structure has a significant impact on RBP binding in the absence of any of these other considerations thus suggesting that internal mRNA secondary structure is an important determinant of RBP binding. My results and evaluation framework also provide a means by which the universality and
predictive power (or lack thereof) of these other possible regulatory mechanisms may be demonstrated. Python scripts to reproduce my analysis, calculate #ATS, assess AUROC and perform my motif finding, as well as my detailed benchmark results, are available in the accompanying CD.

2.5 Methods

Supplementary website http://morrislab.med.utoronto.ca/data/xiao_web/

2.5.1 RBP co-purification and sequence motif data collection

I used RBP co-purification data from six different sources. Though in compiling these data, I made extensive use of matched collections of RBP binding data and consensus sequences compiled by Hogan and colleagues\(^8\) and Ray and colleagues\(^16\). The RBP co-purification data I used was derived from the RNP immunoprecipitation–microarray (RIP-chip) assay. To ensure that I had sufficient statistical power for my analyses, I only used RBPs that co-purified with at least 30 mature mRNA targets. Supplementary File S2.1 in the CD describes the source of each dataset.

Source of transcript sequences The *Drosophila melanogaster* (BDGP5.4), *Saccharomyces cerevisiae* (SGD1.01) and *Homo sapiens* (NCBI36) transcript sequences were downloaded from Ensembl using BioMart (http://www.biomart.org/). For fly and human, I downloaded all cDNA sequences and defined 3’UTRs as the portion of the cDNA downstream of the 3’ end of the coding sequence, as defined by Ensembl. Full-length cDNAs including 5’ and 3’ UTR were not available for most yeast genes, so, like Hogan and colleagues\(^8\), I defined the yeast cDNA as the longest ORF corresponding to each gene plus 200nt upstream and 200nt downstream of the start
and stop codons, respectively, removing any ORF sequence from the upstream or downstream genes.

2.5.2 Defining bound and unbound sets of transcripts

Transcripts were classified into these two sets by comparing their relative enrichment in the RNA fraction co-purifying with the RBP to two thresholds. The positive threshold defined the bound set and the negative threshold defined the unbound set (refer to Figure 2.2 for more details and Supplementary File S2.1 in the CD for the thresholds). I defined relative enrichment using either FDR, Z-score or LOD as reported in the original manuscript. Whenever possible, I used positive and negative thresholds established in the original study describing the data. However, in some cases I used a more permissive negative threshold to increase the statistical power of my analyses.

2.5.3 Quantifying target site accessibility

Target site accessibility was assessed using RNAplfold\(^76\). RNAplfold models co-transcriptional folding by calculating base-pair probabilities using a small window of sequence around the site of interest based on a computational model of thermodynamic stability of RNA secondary structures. Specifically, it estimates the probability that either a binding site, or a single base, is unpaired by calculating local-pair probabilities for bases with a maximal span of \(L\) nucleotides, by sliding a moving window of size \(W\) nucleotides along the input RNA sequence. It computes the probability that a region of \(U\) consecutive nucleotides is unpaired by averaging the probability over all windows of size \(W\) that contain this region. In the experiments, I fixed \(W=80\) and \(L=40\) and set \(U\) to either the width of the consensus sequence or to one when calculated
single-base accessibility. These parameter settings were previously optimized for predicting siRNA binding\textsuperscript{61,77}. When calculating target site accessibilities for a 3’UTR site, I input the entire transcript into RNAplfold to ensure that the target site accessibility for sites immediately downstream of the stop codon incorporated coding sequence.

2.5.4 Scoring accessibility of target sites and their flanking regions

I scored the accessibility of a target site and the flanking region up to \(X\) bases upstream of the site and \(Y\) bases downstream by summing the single base accessibilities (calculated as described above) for the \(X\) upstream and \(Y\) downstream bases and then adding the accessibility of the target site times the length of the target site. I adopted this procedure, rather than calculating the accessibility of the site including its flank, because the latter value often dropped below machine precision (and became inaccurate) for larger values \(X\) and \(Y\).

2.5.5 Statistical tests for the significance of difference between two AUROCs

I used the Delong-Delong-Clarke-Pearson\textsuperscript{136} (DDCP) procedure to assess the significance of differences between the AUROCs on #TS and #ATS. However, because the UcR R package (Lindback) implementation of DDCP that I used not correct for tied-ranks when assessing the significance of difference between two AUROCs, whenever there were tied ranks, I reported the median DDCP \(P\)-value over 100 random permutations of the ordering of these tied ranks.
2.5.6 Motif finding procedure and cross-validation

In general, I used a training set of bound and unbound transcripts to fit motifs which I then assessed using a held-out test set. I applied a two-step strategy to fit motifs. First, I calculated #TS and #ATS AUROCs for all possible 6-mers (including transcripts with no target sites when ranking transcripts, assigning them a score of zero) using the training set. I selected the 6-mers used to seed the next step of motif finding based on these AUROCs. I fit two separate consensus sequence motif models for each RBP, one seeded with the five 6-mers with the highest AUROCs when scored with #ATS, and the other seeded with the five 6-mers with the highest #TS-scored AUROCs. Starting from each seed, I employed an iterative motif-refinement procedure that shortened, lengthened, or introduced degeneracy a single base at a time. At each iteration, the motif that gave the largest AUROC on the training set was selected (measured using #ATS or #TS, as appropriate); the procedure was terminated when the AUROC failed to increase or the associated Bonferroni-corrected Wilcoxon-Mann-Whitney P-value failed to decrease. As with the 6-mers, I also rank transcripts with no target sites when calculating the AUROC. Once the motif finding converged for all five seeds, I selected the motif with the highest AUROC on the training set. I then evaluated the AUROC of each model on the test set to assess its predictive accuracy. I generated 30 training/test set splits using a 3x10 fold cross-validation procedure whereby I randomly split the bound and unbound sets into 10 equally sized bins, trained the motif models on the sequences in nine of the bins and evaluated them on the remaining bin. I repeated this random split three times and collected 30 test set AUROCs for each motif finding method and each transcript scoring method.
2.6 Additional Supporting Data found in the accompanying CD

Supplementary File S2.1. Details of co-purification datasets; previously defined RBP consensus sequences; detailed description of analyzed RBPs; and motif finding results from 3x10 cross-validation tests.

Supplementary File S2.2. Exact $P$-values for all analyses.
Chapter 3

Genome-wide analysis of Staufen-associated mRNAs identifies motifs that confer target specificity

This chapter is derived from the following article: Laver, J.D.*, Li, X.*, Ancevicius, K.*, Morris, Q., Westwood, T., Smibert, C.A., Lipshitz, H. D., Genome-wide analysis of Staufen-associated mRNAs identifies motifs that confer target specificity. NAR, (http://nar.oxfordjournals.org/content/early/2013/08/13/nar.gkt702.full). I performed all of the computational analysis of 3’UTR length, motif finding, as well as structural analysis (described in Results sections 3.3.2-3.3.6). John D. Laver and Kristin Ancevicius performed the RIP-Chip experiments and the relevant work described in results sections 3.3.1. Supplementary Tables 3.1-3.7 can be found in the accompanying CD.
3.1 Abstract

Despite studies that have investigated the interactions of double-stranded RNA-binding proteins like Staufen with RNA in vitro, how they achieve target specificity in vivo remains uncertain. We performed RNA co-immunoprecipitations followed by microarray analysis to identify Staufen-associated mRNAs in early Drosophila embryos. Using computational methods we identified two sequence features that distinguish Staufen’s target transcripts from non-targets. First, these transcripts, as well as those bound by human Staufen1 and 2, have 3’ untranslated regions (UTRs) that are three-to four-fold longer than unbound transcripts. Second, the 3’UTRs of Staufen-bound transcripts are enriched for three specific secondary structures that are highly specific to Staufen bound transcripts. These structures map with high precision to previously identified Staufen-binding regions in Drosophila bicoid and human ARF1 3’UTRs. Our results provide the first systematic, genome-wide analysis showing how a dsRBP achieves target specificity.

3.2 Introduction

RNA-binding proteins (RBPs) direct many co- and post-transcriptional processes. There are a number of different classes of RBPs which are defined by the presence of different RNA-binding domains (RBDs)\(^\text{144}\). One class of RBP is double-stranded RNA binding proteins (dsRBPs), defined by the presence of one or more double-stranded RNA-binding domains (dsRBDs). dsRBDs are characterized by a conserved α β β α fold\(^\text{145-147}\), and bind specifically to double-stranded RNA (dsRNA)\(^\text{148,149}\). Proteins containing dsRBDs have roles in diverse processes, and include E. coli RNase III, Xenopus laevis Xlrbpa, a dsRBP associated with cellular RNAs and ribosomes, the dsRNA-dependent protein kinase PKR, dsRNA-dependent adenosine deaminases (ADARs), and Dicer, an important component of the RNAi machinery.
One of the best-characterized dsRBPs is Staufen, an evolutionarily conserved protein first identified in *Drosophila*\textsuperscript{150}. In *Drosophila* Staufen is essential for localization and translation of oskar mRNA at the posterior of the oocyte\textsuperscript{151-153}, for the anchoring of *bicoid* mRNA in the anterior of the early embryo\textsuperscript{154,155}, and for asymmetric localization of prospero mRNA in dividing embryonic neuroblasts\textsuperscript{156-159}.

Mammals possess two Staufen homologs, Staufen1 and Staufen2, both of which function in developing and adult neurons\textsuperscript{160-163}. Staufen2 has also been shown to segregate asymmetrically during mammalian neural stem cell divisions and to regulate that lineage\textsuperscript{164,165}. Staufen1 has been shown to direct degradation of target RNAs\textsuperscript{166}, to enhance the translation of its targets\textsuperscript{167}, and to regulate pre-mRNA splicing\textsuperscript{168}. A recent report has implicated the *C. elegans* Staufen homolog in exogenous RNAi\textsuperscript{169}.

How dsRBPs like Staufen recognize specific mRNA targets *in vivo* is not well understood. *Drosophila* Staufen contains five dsRBDs, three of which (dsRBD1, dsRBD3 and dsRBD4) bind to dsRNA *in vitro*, and one of which (dsRBD3) binds optimally *in vitro* to a stem-loop containing 12 uninterrupted base pairs (bp)\textsuperscript{170}. However, this stem-loop does not exist within the regions of the *bicoid* 3’UTR to which Staufen binds *in vivo*\textsuperscript{155,171}. In mammals, a 19 bp stem is required for Staufen1 binding to *ARF1* mRNA, its best-characterized target\textsuperscript{172}. However, comparable structures have not been detected in other targets of Staufen1\textsuperscript{166,172,173}.

Intermolecular RNA-RNA interactions may also be important for target recognition by dsRBPs: loop-loop interactions between *bicoid* mRNAs\textsuperscript{171} and interactions between long non-coding (lnc) RNAs and Alu elements in human targets\textsuperscript{174} have both been shown to be important for Staufen binding.

To understand how Staufen recognizes its targets *in vivo*, as well as identify new biological roles for *Drosophila* Staufen, John and Kristin have performed RNA co-immunoprecipitations followed by microarray analysis (RIP-Chip) to identify Staufen mRNA targets in early
Drosophila embryos. They used an anti-GFP antibody to immunoprecipitate transgenic Green Fluorescent Protein (GFP)-tagged Staufen\textsuperscript{158} as well as a synthetic anti-Staufen antibody\textsuperscript{175} to immunoprecipitate endogenous Staufen from wild-type embryos. These experiments identified numerous novel Staufen-associated mRNAs, with a high degree of overlap between the Staufen targets identified by each approach. The functions and localization patterns of these targets support previously known Staufen functions and suggest novel roles for Staufen in early embryos. Using computational methods I identified secondary structures that are enriched among Staufen targets and are highly specific to Staufen-bound transcripts.

3.3 Results

3.3.1 Genome-wide identification of Staufen-associated mRNAs

To identify mRNAs associated with Staufen in early Drosophila embryos, John and Kristin performed RIP-Chip using two complementary approaches. First, John carried out RIP-Chip of endogenous Staufen from wild-type 0-to-3 hr old embryos using a synthetic antibody, anti-Staufen 2A5, that he previously showed immunoprecipitates Staufen protein along with bicoid mRNA\textsuperscript{175}. As a negative control they performed immunoprecipitations using a control antibody (C1) derived from the same synthetic antibody library as anti-Staufen 2A5\textsuperscript{175}. John identified 46 genes whose mRNAs were enriched at least two-fold in Staufen immunoprecipitates compared to negative control immunoprecipitates and had an FDR of less than 5% (Figure 3.1A, B and Table 3.1; Figures 3.2 and 3.3; Supplementary Table S3.1; see Materials and Methods for details). All three previously identified Staufen target mRNAs, bicoid, oskar and prospero, were among these 46, and bicoid mRNA was the most highly enriched target identified. Validation experiments
using reverse-transcription and quantitative PCR (RT-qPCR) are presented in Supplementary Table S3.2.
Figure 3.1. Enrichment of expressed transcripts in Staufen RIPs using synthetic anti-Staufen and anti-GFP Staufen antibodies.
<table>
<thead>
<tr>
<th>Targets identified by synthetic anti-Staufen (FDR ≤5%, fold enrichment ≥ 2) and anti-GFP-Staufen RIP-Chip (FDR ≤5%, fold enrichment ≥ 5)</th>
<th>Targets identified by synthetic anti-Staufen RIP-Chip only (FDR ≤5%, fold enrichment ≥2)</th>
<th>Targets identified by anti-GFP-Staufen RIP-Chip only (FDR ≤5%, fold enrichment ≥5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bicoid</td>
<td>CR14578</td>
<td>CG5830</td>
</tr>
<tr>
<td>dacapo</td>
<td>Phosphoenolpyruvate carboxykinase</td>
<td>Ocelli less</td>
</tr>
<tr>
<td>capping protein beta</td>
<td>Mms19</td>
<td>CG32756</td>
</tr>
<tr>
<td>CR18854</td>
<td>825-Oak</td>
<td>no on or off transient A</td>
</tr>
<tr>
<td>CG4068</td>
<td>CG12519</td>
<td>lethal (3) neo38</td>
</tr>
<tr>
<td>CG32212</td>
<td>sequoia</td>
<td>Roughened</td>
</tr>
<tr>
<td>CG17724</td>
<td>Vacuolar H[+]-ATPase 26kD E subunit</td>
<td>nubbin</td>
</tr>
<tr>
<td>CG3523</td>
<td>Histone demethylase 4B</td>
<td>Flotillin-2</td>
</tr>
<tr>
<td>falafel</td>
<td>Ribosomal protein S27</td>
<td>small wing</td>
</tr>
<tr>
<td>CR32207</td>
<td>split ends</td>
<td>bves</td>
</tr>
<tr>
<td>CG32214</td>
<td>CG4788</td>
<td>CG32767</td>
</tr>
<tr>
<td>Vacuolar H[+] ATPase</td>
<td>virilizer</td>
<td>Sprouty-related protein with</td>
</tr>
<tr>
<td>accessory protein AC45</td>
<td>EVH-1 domain</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>prospero</td>
<td>kugelkern</td>
<td>Tob</td>
</tr>
<tr>
<td>Peroxiredoxin 6005</td>
<td>vielfaltig</td>
<td>CG43736</td>
</tr>
<tr>
<td>Vacuolar H[+] ATPase subunit PPA1-1</td>
<td>CG9977</td>
<td>fusilli</td>
</tr>
<tr>
<td>partner of paired</td>
<td>Neurofibromin 1</td>
<td>CG33932</td>
</tr>
<tr>
<td>oskar</td>
<td>CG14915</td>
<td>pasilla</td>
</tr>
<tr>
<td>CG14100</td>
<td>Mediator complex subunit 14</td>
<td>punt</td>
</tr>
<tr>
<td>Protein phosphatase 19C</td>
<td>CG32267</td>
<td>CG10777</td>
</tr>
<tr>
<td>CG18273</td>
<td>CG17270</td>
<td>CG5966</td>
</tr>
<tr>
<td>staufen</td>
<td>CR18166</td>
<td>cAMP-dependent protein kinase R1</td>
</tr>
<tr>
<td>Dorsal switch protein 1</td>
<td>vein</td>
<td></td>
</tr>
<tr>
<td>squeeze</td>
<td>TNF-receptor-associated factor 6</td>
<td></td>
</tr>
<tr>
<td>Polycomb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG31688</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1. Staufen-associated mRNAs identified by synthetic anti-Staufen RIP-Chip (fold-enrichment ≥ 2) and anti-GFP-Staufen RIP-Chip (fold-enrichment ≥ 5).
Figure 3.2. Microarray expression of transcripts from w^{1118} embryo extract used as input for synthetic anti-Staufen RIPs.

(A) For the three biological replicates assayed, the log₂ signal intensities of the 16,623 transcripts probed on the arrays were plotted to allow replicate-to-replicate comparisons. The square of the Pearson correlation coefficient was calculated for each pair of replicates and was used to determine the degree of similarity between them. (B) The distribution of transcript expression was plotted using the average log₂ signal intensity of each transcript. The mean log₂ signal intensity across all transcripts (8.42), which was used as the expression threshold for SAM one-class analysis, is denoted by a dotted line.
Figure 3.3. Replicate-to-replicate comparisons of transcript microarray signal intensities from the immunoprecipitated samples of the synthetic anti-Staufen and control synthetic antibody RIPs.

For the three biological replicates assayed, the log₂ signal intensities of the 16,623 transcripts probed on the arrays were plotted to allow replicate-to-replicate comparisons of (A) the synthetic anti-Staufen RIP samples and (B) the control RIP samples. The square of the Pearson correlation coefficient was calculated for each pair of replicates and was used to determine the degree of similarity between them.
To complement the synthetic antibody RIP-Chip, Kristin also performed RIP-Chip using flies expressing GFP-tagged Staufen\textsuperscript{158} immunoprecipitated with a commercially available anti-GFP antibody. Western blotting showed that this antibody successfully immunoprecipitated the fusion protein from transgenic embryo extract (Figure 3.4A). This enabled her to identify 503 genes whose mRNAs were enriched at least two-fold in the anti-GFP immunoprecipitates compared to an anti-FLAG control, and had an FDR of less than 5% (Figure 3.1C, D; Figures 3.5 and 3.6; Supplementary Table S3.3). As with the synthetic antibody, all three previously known Staufen mRNA targets were identified and \textit{bicoid} was again the most highly enriched target. Validation experiments are presented in Supplementary Table S3.2.
Figure 3.4. Western blots examining Staufen immunoprecipitation recovery and Staufen expression levels.

(A) GFP-Staufen embryo extract was used to perform an immunoprecipitation targeting the tagged Staufen. Western blotting with anti-GFP shows that in the input and anti-GFP immunoprecipitated samples, but not the control, a prominent band is seen near 175kDa, which represents GFP-Staufen. (B) The expression of both endogenous and transgenic GFP-tagged Staufen was assayed across three different Drosophila lines. A single band representing endogenous Staufen can be seen in the w^{1118} lane while this band, along with a higher molecular weight band representing GFP-Staufen, can be seen in the transgenic GFP311 lane. Both bands are absent from the lane containing extract from embryos produced by staufen mutant females (stau^{P93}/stau^{P93}). After normalizing to tubulin, the band corresponding to GFP-Staufen was 1.8-1.9 times as intense as the endogenous Staufen band in the w^{1118} lane.
Figure 3.5. Microarray expression of transcripts from GFP-Staufen embryo extract used as input for anti-GFP-Staufen RIPs.

(A) For the three biological replicates assayed, the log₂ signal intensities of the 16,623 transcripts probed on the arrays were plotted to allow replicate-to-replicate comparisons. The square of the Pearson correlation coefficient was calculated for each pair of replicates and used to determine the degree of similarity between them. (B) The distribution of transcript expression was plotted using the average log₂ signal intensity of each transcript. The mean log₂ signal intensity across all transcripts (8.02), which was used as the expression threshold for SAM one-class analysis, is denoted by a dotted line.
Figure 3.6. Replicate-to-replicate comparisons of transcript microarray signal intensities from the samples of the anti-GFP and control anti-FLAG RIPs.

For the three biological replicates assayed, the log$_2$ signal intensities of the 16,623 transcripts probed on the arrays were plotted to allow replicate-to-replicate comparisons of (A) the anti-GFP RIP samples and (B) the anti-FLAG control RIP samples. The square of the Pearson correlation coefficient was calculated for each pair of replicates and was used to determine the degree of similarity between them.
41 of the 46 genes (89%) identified as Staufen targets using the synthetic antibody were also identified by the anti-GFP RIP-Chip (Fisher’s exact test, $P < 3 \times 10^{-16}$; Figure 3.7A). Moreover, 36 out of the 46 targets identified by the synthetic antibody lay within the 100 most highly enriched targets in the anti-GFP experiment (Figure 3.7B). Therefore, there is a high degree of similarity between the results of the two experiments.

However, the anti-GFP RIP-Chip identified 10-fold more targets than the synthetic antibody RIP-Chip. This discrepancy is likely attributable to differences both in antibody affinities and in expression levels of endogenous versus transgenic Staufen. For example, RIPs using anti-GFP antibody from GFP-Staufen extract gave a three-fold higher enrichment of bicoid mRNA than synthetic anti-Staufen antibody from the same extract (Figure 3.7C), most likely as a result of more efficient pull-down of Staufen by the anti-GFP antibody. However, RIPs using the synthetic antibody from GFP-Staufen extract yielded 1.5- to 2-fold higher enrichment of bicoid mRNA than RIPs using the same antibody from wild-type extract (Figure 3.7C). Since GFP-Staufen is present at 1.5- to 2.0-fold higher levels in GFP-Staufen extract than endogenous Staufen in wild-type extract (Figure 3.4B), overexpression of GFP-Staufen also appears to contribute to binding to a larger number of mRNAs.

This raises the possibility that a subset of the 503 targets identified by the anti-GFP RIP-Chip may not be natural Staufen targets, and may only be bound in the context of its overexpression. We therefore created a high-confidence list of 48 targets by applying a more stringent, five-fold enrichment cut-off to the anti-GFP data (Figure 3.1E, F; Supplementary Table S3.3). These included bicoid, oskar and prospero, and shared a total of 25 transcripts with the synthetic-antibody-identified targets (Figure 3.7A, B) (Fisher’s exact test, $P < 3 \times 10^{-16}$).

In order not to exclude low-affinity but real targets of Staufen identified in the anti-GFP RIP-Chip with a two-fold enrichment cut-off, but also to ensure particular consideration of the high-confidence targets identified using the synthetic antibody and the anti-GFP RIP-Chip with a five-
fold cut-off, the subsequent analyses were conducted on all three lists of Staufen-associated transcripts.
Figure 2

A

B

C

Synthetic anti-Staufen
Anti-GFP 5-fold
Anti-GFP 2-fold

503 genes

1.49
15.05
1.22
25.44
1.99
78.45

0
10
20
30
40
50
60
70
80
90

Fold enrichment

Wild-type extract
GFP-Staufen extract

Rpl32
dbd

Rpl32
dbd

Rpl32
dbd

Synthetic anti-Staufen
anti-GFP

5-fold enrichment

2-fold enrichment
Figure 3.7. Comparison of the synthetic anti-Staufen and anti-GFP-Staufen RIPs.

(A) Venn diagram (generated using the BioVenn web application (80)) shows overlap of Staufen targets from the synthetic anti-Staufen RIP with a fold enrichment cut-off of at least two (blue) and the transgenic anti-GFP RIPs with fold enrichment cut-offs of at least two (red) or five (yellow). (B) The 503 genes from the anti-GFP two-fold list were ranked according to decreasing fold enrichment and the 41 overlapping genes from the synthetic anti-Staufen two-fold list were then overlaid in black showing that they represent genes with some of the highest fold-enrichments. (C) RT-qPCR analysis of the enrichment of the target mRNA bicoid and the reference mRNA RpL32 in Staufen RIPs conducted using wild-type extract and synthetic anti-Staufen, transgenic GFP-Staufen extract and synthetic anti-Staufen, and transgenic GFP-Staufen extract and anti-GFP. Each bar represents the average fold enrichment of the respective transcript in the anti-Staufen RIPs relative to the appropriate control. Error bars represent the standard error of the mean for n=3 biological replicates.
3.3.2  *Drosophila* and human Staufen targets have unusually long 3'UTRs

I next assessed features of the transcript sequence that might distinguish Staufen-associated mRNAs from co-expressed, non-target mRNAs. First, I compared 5'UTR, open reading frame (ORF), and 3'UTR lengths and found that the median 3’UTR length of targets was three- to four-fold greater than that of the co-expressed non-targets (hereafter “non-targets”) for all three lists with a Bonferroni corrected Wilcoxon-Mann-Whitney (WMW) rank sum of $P < 10^{-4}$ (Figure 3.8A; Supplementary Table S3.4). The 5’UTRs of targets on both anti-GFP lists were also significantly longer than non-targets, although to a lesser extent (two- to three-fold; Bonferroni corrected WMW $P < 10^{-5}$) and this was not the case for the synthetic anti-Staufen targets (fold = 1.78; Bonferroni-corrected WMW $P = 0.11$). The median length of target ORFs on the anti-GFP two-fold list was marginally longer (Bonferroni corrected WMW $P = 0.04$) but this was not true for the other two lists.

I next re-analyzed three previously published RIP-Chip experiments identifying human Staufen1 and 2 targets\(^{173}\) and found that they also had approximately four-fold longer 3’UTRs than co-expressed non-targets (Bonferroni corrected WMW $P < 10^{-7}$) (Figure 3.8B and Supplementary Table S3.4; see Materials and Methods). There was no significant difference in 5’UTR or ORF length for these targets (Figure 3.8B and Supplementary Table S3.4).

In summary, 3’UTR length is a major feature that distinguishes both the *Drosophila* and human Staufen targets from non-targets. To determine whether long 3’UTRs were unique to Staufen targets, we asked whether targets of several single-stranded RNA-binding proteins also have long 3’UTRs. We calculated the ratios of median 3’UTR lengths of single-stranded RBP targets to the median lengths of the 3’UTRs of co-expressed non-targets identified using RIP-Chip for *Drosophila* PUM\(^{50}\), human ELAVL1\(^{109}\), human PTB\(^{176}\) and human PUM1\(^{135}\). In all cases, the
mRNAs in these target sets also had long 3’UTRs; however the fold-increase was less than for Staufen targets: 1.5- to 2.5-fold rather 3.0- to 4.5-fold (Supplementary Table S3.4 and Figure 3.9).

Thus, while Staufen targets do have particularly long 3’UTRs, long 3’UTRs are a feature of the targets of both double-stranded and single-stranded RBPs. Indeed, long 3’UTRs may be a feature of post-transcriptionally regulated transcripts in general, particularly maternal mRNAs and nervous-system isoforms of mRNAs.
Figure 3.8. *Drosophila* and human Staufen targets have unusually long 3'UTRs.

(A) Bar plots showing the ratios of the median length of the 5'UTR, ORF, and 3'UTR of Drosophila Staufen targets to the median length of the co-expressed non-targets. (B) The ratios of median lengths of human Staufen target transcripts to the median lengths of the co-expressed non-targets. The human Staufen1 targets were identified using anti-HA RIP-Chip from HEK293T cells transfected with Stau155-HA expressor (with five-fold enrichment cut-off, FDR < 5%). The human Staufen2 targets were identified using anti-HA RIP-Chip from HEK293T cells transfected with Stau259-HA expressor or Stau262-HA expressor (with five-fold enrichment cut-off, FDR < 5%). The statistical significance of the differences between the lengths of targets and non-targets was assessed using Bonferroni-corrected Wilcoxon-Mann-Whitney (WMW) P-values: *: $P < 0.05$, **: $P < 0.001$, ***: $P < 10^{-6}$. Exact numbers are given in Supplementary Table S3.4.
Figure 3.9. Single-stranded RBP targets have long 3’UTRs.

The ratios of median 3’UTR lengths of single-stranded RBP targets to the median lengths of the 3’UTRs of co-expressed non-targets identified for Drosophila PUM [177], human ELAVL1 [178], human PTB [179] and human PUM1 [135] using RIP-Chip. The statistical significance of the differences between the lengths of targets and non-targets was assessed using Bonferroni-corrected Wilcoxon-Mann-Whitney (WMW) P values: *: P < 0.05, **: P < 0.001, ***: P < 10^{-6}. Exact numbers are given in Supplemental Table S3.4.
3.3.3 High-confidence *Drosophila* Staufen target mRNAs are enriched for specific structural motifs in their 3’UTRs

We next asked whether we could identify specific structural motifs in the 3’UTRs of Staufen targets that would distinguish them from co-expressed non-targets. Specifically, we searched for double-stranded stems of varying length, ranging from 1 to 22 bp, and with varying degrees of imperfect pairing, and asked whether any such structures were enriched among Staufen targets compared to non-targets. Local folding of mRNA is a better predictor of its secondary structure and protein interaction than global folding\(^\text{82}\); but local folding tools, like RNAplfold, only output the probability that a single base is paired. As such, we adopted a hybrid strategy to identify the characteristics of stems bound by Staufen: we first identified regions of the mRNA sequence likely to be paired; then we assessed Staufen’s preferences for paired regions of various sizes (1 to 22nts) and, finally, filtered the preferred regions for those that were actually stems. Figure 3.10 diagrams the steps in this process.
Figure 3.10. Schematic of the *in silico* assays for discovery of Staufen’s binding preferences.

On the leftmost structures, red represents a 15 of 19 motif while black indicates partners of this motif (see Materials and Methods and Results for details).
To identify 3’UTR regions likely to be in stems, for all \( M \)-mers (\( M = 1 \) to 22 continuous nucleotides) we estimated the probability that a number, \( N \), of these \( M \) continuous bases were paired using RNAplfold. This is a necessary but not sufficient condition for the \( M \)-mer to be in a stem with \( N \) of its bases paired. For each \( M \)-mer, we set a range for the allowed number of mismatched or unpaired bases from zero to either four or \( M \) divided by four, whichever was less (\( e.g., \) where \( M-N = 0 \) to min[4,\( M/4 \)]). As an example, based on this definition, an \( M \)-mer of 19 with from 15 to 19 paired bases (where the lowest number of paired bases is designated \( N \)) is given the designation “15 of 19” (\( i.e., \) \( N \) of \( M \)). Also note that we did not allow unpaired bases at the first and last position of the \( M \)-mer. We then scanned each 3’UTR for each \( N \) of \( M \) combination, and scored a region as containing a particular \( N \) of \( M \) “hit” if the probability that the region contained \( N \) of \( M \) paired bases was found in the top 1% of the \( N \) of \( M \) probabilities across all 3’UTRs (see Supplementary Table S3.5 for the cutoffs). We refer to each \( N \) of \( M \) as a different “motif” because each corresponds to a contiguous sequence of bases, and we call the hits “motif matches”. Figure 3.10B diagrams several motif hits. We assigned each transcript an \( N \) of \( M \) “motif hit score” equal to the sum of the probabilities of all hits in that transcript’s 3’UTR (see Materials and Methods for details) and asked whether these scores were significantly higher in the Staufen target transcripts versus the non-targets using a one-sided Wilcoxon rank sum test. In addition, to ensure that any significant increase in \( N \) of \( M \) motif scores of Staufen targets was not solely a result of the differences in 3’UTR length, we set the Rank sum \( P \) value of the 3’UTR length as the baseline. We use AUROC to indicate the effect size of the enrichment as it has a linear relationship with the Wilcoxon rank sum test statistic; in vast majority of cases, the feature (\( i.e., \) \( N \) of \( M \) motif score or 3’ UTR length) associated with a higher AUROC on a dataset also has a more significant \( P \) value. AUROC is also a measure of classification accuracy; in this context, if a feature has a higher AUROC, its scores are better predictors of Staufen binding in these data.
We performed this analysis on each of the three sets of Drosophila Staufen targets described above as well as, for comparison, Pumilio targets. Analysis of the Staufen targets identified by anti-GFP RIP-Chip with a two-fold enrichment cut-off and of the Pumilio targets did not reveal any N of M motifs with higher AUROC than 3’UTR length (Supplementary Table S3.6). In contrast, analyses of the targets identified using either synthetic Staufen or anti-GFP Staufen with five-fold cut-off revealed AUROCs above baseline for several N of M motifs (Figure 3.11; Supplementary Table S3.6): for the synthetic antibody targets, these included values of M ranging from 4 to 21, with the highest peak at the 10 of 12 motif (Figure 3.11A, Supplementary Table S3.6); for the anti-GFP five-fold targets, these included values of M ranging from 8 to 22, with the highest peak at the 16 of 19 motif (Figure 3.11C, Supplementary Table S3.6).

Many of the motifs identified above differed by only one in M or in N, raising the possibility that some of the motifs performed well simply because they were imperfect predictors of the presence of a related motif. Indeed, the values of the transcript scores for similar motifs were extremely highly correlated (Figure 3.12). Therefore, to select a core, non-redundant set of N of M motifs that collectively explained Staufen’s binding preferences, we performed LASSO logistic regression on the synthetic antibody and anti-GFP five-fold lists. As potential features to distinguish Staufen targets from non-targets, we used 3’UTR length and all of the N of M motif scores identified by us above as individually having AUROCs above baseline. This sparse regression analysis assigns weights to each feature, with the most important features being assigned the greatest weights and the least important features receiving weights of zero. For both data sets, LASSO assigned non-zero weights to 10 of 12 (which had the greatest weights: 0.34 and 0.43), and to 15 of 19 (which had a weight of 0.13 for both target sets) (Figure 3.11B, D). Non-zero weights were also given to 9 of 10, 11 of 11, 14 of 16, and 18 of 20. The logistic regression models containing these collections of motifs had a significantly better fit to the Staufen-binding data from their corresponding sets than ones based on 3’UTR length alone (likelihood ratio test $P < 10^{-7}$ for synthetic antibody and $P < 10^{-11}$ for the anti-GFP, five-fold;
Figure 3.11; Supplementary Table S3.6), indicating that the presence of these motifs explains Staufen’s binding preferences better than the strong bias in 3’UTR length noted earlier.
Figure 3.11. Specific double-stranded structures are enriched in Staufen target transcript 3'UTRs.

(A, C) Wilcoxon rank sum \( P \) values were used to assess how well a particular double-stranded stem could distinguish between Staufen targets and co-expressed non-targets that were defined by (A) synthetic antibody RIP-Chip (with
two-fold enrichment cut-off, FDR ≤ 5%) or (C) anti-GFP RIP-Chip (with five-fold enrichment cut-off, FDR ≤ 5%). The test was performed on stems of varying length ranging from 1 to 22 base pairs (indicated on the X-axis) and with varying degrees of imperfect pairing ranging from 0 to 4 mismatches (indicated by the different colors and symbols). Among these, only the stems with better more significant P value than for 3’UTR length (the baseline) are shown and were tested in the analysis shown in (B) and (D). (B, D) Using the stems identified in (A) and (B) as well as the 3’UTR length as the features, LASSO regression was trained to select the features most relevant to Staufen target prediction (i.e., features with non-zero weights). Compared to the training model using 3’UTR length only, the LASSO-selected features significantly improved the prediction of Staufen binding: likelihood ratio test $P < 10^{-7}$ (B) and $P < 10^{-11}$ (D). Exact numbers are given in Supplementary Table S3.6.
Figure 3.12. Correlations between the scores of the double-stranded motifs searched for in *Drosophila* 3'UTRs.
Having identified this core set of motifs through our analysis of 3’UTRs, we also asked whether there was enrichment for these motifs in the 5’UTRs or ORFs of Staufen-bound mRNAs. To do this, we asked whether the motif scores computed for the 5’UTR or ORF had significantly higher AUROCs than just the length of the corresponding region, at the task of classifying transcripts according to Staufen binding on either the synthetic antibody or anti-GFP five-fold target sets (Supplementary Table S16). For both 5’UTR and ORF, none of the motifs was a better predictor than length. In addition, notably, the motif hits in the 3’UTR had consistently higher AUROCs than those in either the ORF or 5’UTR (e.g., AUROCs of 0.73 to 0.90 for those in the 3’UTR versus 0.48 to 0.63 for 5’UTR and ORF; see Table 3.2).
A. Synthetic anti-Staufen

<table>
<thead>
<tr>
<th>Motif</th>
<th>5'UTR</th>
<th></th>
<th>ORF</th>
<th></th>
<th>3'UTR</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>WMW $P$ value</td>
<td>AUROC</td>
<td>WMW $P$ value</td>
<td>AUROC</td>
<td>WMW $P$ value</td>
<td>AUROC</td>
</tr>
<tr>
<td></td>
<td>1.73E-03</td>
<td>0.65</td>
<td>0.92</td>
<td>0.49</td>
<td>4.68E-07</td>
<td>0.74</td>
</tr>
<tr>
<td>15 of 19</td>
<td>1.89</td>
<td>0.50</td>
<td>0.63</td>
<td>0.55</td>
<td>4.55E-08</td>
<td>0.75</td>
</tr>
<tr>
<td>10 of 12</td>
<td>0.68</td>
<td>0.54</td>
<td>0.88</td>
<td>0.54</td>
<td>5.78E-11</td>
<td>0.80</td>
</tr>
<tr>
<td>14 of 16</td>
<td>1.09</td>
<td>0.53</td>
<td>1.18</td>
<td>0.53</td>
<td>1.33E-09</td>
<td>0.77</td>
</tr>
<tr>
<td>9 of 10</td>
<td>0.34</td>
<td>0.56</td>
<td>0.81</td>
<td>0.54</td>
<td>1.89E-09</td>
<td>0.78</td>
</tr>
<tr>
<td>11 of 11</td>
<td>0.33</td>
<td>0.56</td>
<td>1.86</td>
<td>0.50</td>
<td>9.01E-07</td>
<td>0.73</td>
</tr>
<tr>
<td>18 of 20</td>
<td>0.47</td>
<td>0.55</td>
<td>1.51</td>
<td>0.48</td>
<td>7.06E-08</td>
<td>0.74</td>
</tr>
</tbody>
</table>

B. Anti-GFP, 5-fold

<table>
<thead>
<tr>
<th>Motif</th>
<th>5'UTR</th>
<th></th>
<th>ORF</th>
<th></th>
<th>3'UTR</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>WMW $P$ value</td>
<td>AUROC</td>
<td>WMW $P$ value</td>
<td>AUROC</td>
<td>WMW $P$ value</td>
<td>AUROC</td>
</tr>
<tr>
<td></td>
<td>1.96E-07</td>
<td>0.75</td>
<td>0.50</td>
<td>0.53</td>
<td>5.70E-14</td>
<td>0.85</td>
</tr>
<tr>
<td>15 of 19</td>
<td>0.01</td>
<td>0.61</td>
<td>0.04</td>
<td>0.61</td>
<td>2.83E-18</td>
<td>0.88</td>
</tr>
<tr>
<td>10 of 12</td>
<td>0.03</td>
<td>0.61</td>
<td>0.08</td>
<td>0.60</td>
<td>2.02E-18</td>
<td>0.89</td>
</tr>
<tr>
<td>14 of 16</td>
<td>0.05</td>
<td>0.60</td>
<td>0.14</td>
<td>0.58</td>
<td>3.21E-20</td>
<td>0.90</td>
</tr>
<tr>
<td>9 of 10</td>
<td>0.02</td>
<td>0.61</td>
<td>0.13</td>
<td>0.59</td>
<td>2.56E-17</td>
<td>0.88</td>
</tr>
<tr>
<td>11 of 11</td>
<td>4.64E-03</td>
<td>0.63</td>
<td>0.29</td>
<td>0.57</td>
<td>1.91E-17</td>
<td>0.88</td>
</tr>
<tr>
<td>18 of 20</td>
<td>0.01</td>
<td>0.61</td>
<td>0.43</td>
<td>0.56</td>
<td>2.24E-20</td>
<td>0.89</td>
</tr>
<tr>
<td>Motif</td>
<td>5'UTR WMW P value</td>
<td>5'UTR AUROC</td>
<td>ORF WMW P value</td>
<td>ORF AUROC</td>
<td>3'UTR WMW P value</td>
<td>3'UTR AUROC</td>
</tr>
<tr>
<td>---------</td>
<td>------------------</td>
<td>------------</td>
<td>----------------</td>
<td>-----------</td>
<td>------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Length</td>
<td>6.82E-27</td>
<td>0.67</td>
<td>9.47E-04</td>
<td>0.55</td>
<td>1.43E-57</td>
<td>0.75</td>
</tr>
<tr>
<td>15 of 19</td>
<td>1.91E-11</td>
<td>0.60</td>
<td>8.48E-08</td>
<td>0.59</td>
<td>8.79E-38</td>
<td>0.69</td>
</tr>
<tr>
<td>10 of 12</td>
<td>1.19E-10</td>
<td>0.60</td>
<td>2.05E-06</td>
<td>0.58</td>
<td>3.85E-40</td>
<td>0.70</td>
</tr>
<tr>
<td>14 of 16</td>
<td>9.33E-11</td>
<td>0.59</td>
<td>2.42E-07</td>
<td>0.58</td>
<td>3.34E-46</td>
<td>0.71</td>
</tr>
<tr>
<td>9 of 10</td>
<td>1.42E-09</td>
<td>0.59</td>
<td>1.77E-06</td>
<td>0.58</td>
<td>2.42E-39</td>
<td>0.70</td>
</tr>
<tr>
<td>11 of 11</td>
<td>4.42E-13</td>
<td>0.61</td>
<td>6.30E-05</td>
<td>0.56</td>
<td>2.11E-43</td>
<td>0.71</td>
</tr>
<tr>
<td>18 of 20</td>
<td>4.13E-13</td>
<td>0.60</td>
<td>1.13E-04</td>
<td>0.56</td>
<td>2.63E-47</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 3.2. Search results of the six computationally identified motifs in Drosophila 5'UTRs, ORFs and 3'UTRs.
Having data strongly suggesting that Staufen binds 3’UTRs and having identified several potential motifs that are enriched in the 3’UTRs of Staufen’s targets, we combined the Staufen-associated transcripts from the synthetic antibody and GFP five-fold data sets in order to further refine our model of Staufen binding. This created a new set of Staufen target transcripts consisting of the union of the Staufen targets from the two datasets, and a new set of co-expressed non-target consisting of the intersection of the non-targets from the two datasets. Repeating our regression analysis on these new sets, we found that the fit of the model containing only the 10 of 12 and 15 of 19 motifs (and 3’UTR length) was statistically indistinguishable from one that included all six motifs (and 3’UTR length) (P > 0.05, one-tailed LRT). Furthermore, the model containing only the 10 of 12 and 15 of 19 motifs had a significantly better fit than one that only contained one of these two motifs (P < 0.05 for both, one-tailed LRT, Bonferroni-corrected). As such, for further analysis, we considered only 10 of 12 and 15 of 19.

### 3.3.4 Computational analysis of the properties of dsRNA stems bound by Staufen

The above analysis provides information for only one side of a potential dsRNA structure (i.e., the red strands illustrated in Figure 3.10B). As such, a single N of M designation describes a large number of potential secondary structures (for some examples see the combinations of red and black strands in Figure 3.10B). Thus to further refine our analysis we next sought to identify properties of the dsRNA stems that corresponded to the 10 of 12 and 15 of 19 motif hits. To do so, we used Sfold to predict the secondary structure of a region of approximately 300 nucleotides centered on each of the 10 of 12 and 15 of 19 motif matches. Sfold outputs two results, both of which we considered in our analysis: a set of 1,000 structures sampled from the ensemble of all possible secondary structures, and the “centroid” structure, which is the single structure with the highest total agreement with all 1,000 samples. Using the Sfold results, we
then identified those centroids where the motif hit satisfied the following three criteria, thus placing it in a dsRNA stem: i) at least $N$ of its $M$ bases had to be paired, including the first and last bases; ii) its “partner region”, which is the transcript sequence between the bases that pair with the first and last bases of the $N$ of $M$ hit, had to pair only with bases in the hit (i.e., contain no hairpins); iii) the motif hit had to pair only with bases in its partner region (compare Figure 3.10B and C to see examples of structures that are filtered out, or retained, using these criteria). Based on these analyses we found that a significantly higher percentage of motif hits were in dsRNA stems in the Staufen targets versus non-targets (Table 3.3, $P < 10^{-9}$). We then removed hits that did not correspond to dsRNA stems and combined hits that were on either side of the same dsRNA stem into a single structure consisting of the motif hit and the sequence with which it pairs. We call these structures $[12,10]$ and $[19,15]$ where the first number indicates that at least one side of the dsRNA stem spans exactly 12 (or 19) nucleotides and the second number indicates that at least 10 (or 15) bases are paired on that side (For examples, see Figure 3.10C). Note that the partner region could, in principle, span a different number of nucleotides and have fewer or more unpaired bases.
### Table 3.3: Sfold validation of 10 of 12 and 15 of 19 motif hits.

<table>
<thead>
<tr>
<th></th>
<th>Staufen targets</th>
<th>Co-expressed non-targets</th>
<th>Hypergeometric P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 of 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of motif hits in dsRNA stem</td>
<td>672</td>
<td>3272</td>
<td>$1.83 \times 10^{-10}$</td>
</tr>
<tr>
<td># of motif hits</td>
<td>954</td>
<td>5469</td>
<td></td>
</tr>
<tr>
<td>% motif hits in dsRNA stem</td>
<td>70.44</td>
<td>59.83</td>
<td></td>
</tr>
<tr>
<td># of genes with at least one hit in dsRNA stem</td>
<td>45</td>
<td>572</td>
<td>$4.41 \times 10^{-11}$</td>
</tr>
<tr>
<td># of genes with at least one hit</td>
<td>55</td>
<td>1528</td>
<td></td>
</tr>
<tr>
<td>% genes with at least one hit in dsRNA stem</td>
<td>81.82</td>
<td>37.43</td>
<td></td>
</tr>
</tbody>
</table>

|          |                 |                          |                        |
|----------|-----------------|--------------------------|                        |
| 15 of 19 |                 |                          |                        |
| # of motif hits in dsRNA stem | 487 | 1626 | $1.71 \times 10^{-36}$ |
| # of motif hits | 954 | 5469 |
| % motif hits in dsRNA stem | 51.05 | 29.73 |
| # of genes with at least one hit in dsRNA stem | 36  | 344  | $3.54 \times 10^{-11}$ |
| # of genes with at least one hit | 55  | 1528 |
| % genes with at least one hit in dsRNA stem | 65.45 | 22.51 |
Upon further analysis of the [12,10] and the [19,15] structures, we found clear differences between the Staufen targets and non-targets for both structures. Specifically, we assessed the presence of various structural features: (i) unpaired bases, which refer to those bases that do not have a corresponding partner base on the other strand of a stem, (ii) mismatches, which refer to two bases that are found across from one another in a secondary structure but are not canonical base pairs, (iii) internal loops, which are loops emanating from within one strand of the dsRNA stem that contain at least one mismatch and, possibly, some unpaired bases, and (iv) bulge loops, which are loops emanating from within one strand of the stem and that contain only one or more unpaired bases. These are diagrammed in Figure 3.10A.

Comparing Staufen targets to non-targets, our analysis revealed that the [12,10] and [19,15] structures in the targets had significantly fewer unpaired bases (Supplementary Figure S8) as well as a weak preference against mismatches. The bias against unpaired bases also manifested as a significant depletion for bulges in the positive set (Figure 3.13). In addition, there was a significant preference for small internal loops or bulges in the Staufen targets (Figure 3.14). Taken together, these results suggest an overall model in which Staufen binds [12,10] and [19,15] dsRNA stems that have a small number of mismatches, zero or few unpaired bases, and short internal loops.
Figure 3.13. Characteristics of the stems bound by Staufen at the level of individual sites.

The significance was evaluated by Wilcoxon rank sum $P$ values.
Figure 3.14. Loop size in positive and negative sets.

The significance was evaluated by Wilcoxon rank sum $P$ values.
Many transcripts in both the target and non-target sets contain multiple $[12,10]$ and $[19,15]$ structures, as such, we then considered the same criteria at the transcript level to determine features associated with the “best” Staufen site in the transcript by assigning each transcript a mismatch, unpaired base, loop count or loop length equal to the minimum of those values across all $[12,10]$ and $[19,15]$ structures in the transcript’s 3’UTR. This analysis also revealed a significantly lower number of unpaired bases in the “best” structure in Staufen target transcripts compared to non-targets, which also manifested as an even stronger preference for balanced stems (i.e., those containing only paired or mismatched bases and no unpaired bases (Figure 3.15). In particular, 82% of Staufen targets that had at least one $[12,10]$ structure contained a balanced $[12,10]$ structure (compared with 47% in non-targets). We use the designation $[12,10,0]$ to describe these balanced structures where the last number refers to the maximum number of unpaired bases in the structure. The enrichment among the Staufen targets for transcripts containing balanced $[19,15]$ structures (i.e., $[19,15,0]$) was even more striking (67% versus 12%). Because many more Staufen target transcripts have $[19,15]$ structures, if we define a Staufen-bound transcript as one containing a $[19,15]$ that is balanced (i.e., $[19,15,0]$), this rule would be satisfied by 44% (n=24) of the positive transcripts but only 3% of the negative transcripts (n=42).
Figure 3.15. Characteristics of the stems bound by Staufen.

We compared the 10 of 12 (A) and 15 of 19 (B) structures in the Staufen targets versus non-targets, by characterizing the structural features: i) number of mismatches, ii) number of unpaired bases, iii) number of bulges, iv) number of internal loops. If there was more than one structure in a transcript’s 3’UTR, the feature with the minimal score was used to represent the gene. The Wilcoxon rank sum test was used to assess the feature in Staufen targets (the blue cdfplot) versus non-targets (the red cdfplot).
We note that Staufen sites defined by previous structural, in vitro and/or in vivo studies are not nearly as predictive of its binding as the $[12,10]$ and $[19,15]$ structures: using the criteria described above, only 31% of all positives and 4% of the negatives contain a perfect 12bp region of dsRNA (i.e., $[12,12,0]$), which was defined as an optimal structure for Staufen dsRBD3 binding and none of the positive transcripts and only two of the negative transcripts (0.1%) contain a perfect 19bp region of dsRNA (i.e., $[19,19,0]$) defined as an optimal site for human Staufen binding to ARF1’s 3’UTR.

That said, the $[19,15,0]$ rule does not explain all instances of Staufen binding; in particular, the bicoid 3’UTR does not contain one of these high-confidence sites. However, we can slightly refine our model of the $[12,10]$ and $[19,15]$ structures to $[12,10,2]$ and $[19,15,4]$, respectively (Figure 3.10). A model including these latter structures is better at distinguishing Staufen targets from non-targets than the $[12,10]$ and $[19,15]$ structures. While this model removes individual sites from the Staufen target set, all Staufen targets with a $[12,10]$ or $[19,15]$ structure contain at least one $[12,10,2]$ or $[19,15,4]$ structure, respectively. On the other hand, this rule reduces the number of non-target transcripts with these structures by removing those that have a very large number of unpaired bases in the partner region (e.g., one $[19,15]$ structure in the non-target set had >100 unpaired bases). Note that all $[19,15,0]$ structures are also $[19,15,4]$ structures because the latter includes structures with three, two, one and zero unpaired bases. In total, 65% of Staufen targets and 18% of non-targets contain a $[19,15,4]$ structure while 82% of Staufen targets and 33% of non-targets contain a $[12,10,2]$ structure.

3.3.5 Definition of Staufen-recognized structures (SRSs)

Having defined the structural characteristics of the dsRNA stems bound by Staufen, we next mapped these structures onto: i) the 3’UTRs of the Staufen target transcripts; ii) length-matched
non-target transcript 3’UTRs; and iii) a random subset of non-target 3’UTRs (Figure 3.16). This led us to note a substantial overlap of \([19,15,4]\) and \([12,10,2]\) structures in the Staufen target set. As such, we sought to determine whether the \([12,10,2]\) structures independently are associated with Staufen binding or if they are simply features of more predictive \([19,15,4]\) structures. To do so, we first assessed whether the presence of \([12,10,2]\) in the absence of \([19,15,4]\) was predictive of Staufen binding by comparing Staufen target and non-target transcripts that did not contain a \([19,15,4]\) structure. We found that 47% (9 of 19) of such transcripts in the Staufen target set have a \([12,10,2]\) structure, whereas only 22% (279 of 1245) of the non-targets have one \((P = 0.02, \text{Hypergeometric test})\). We then asked whether Staufen target transcripts were more likely to have a \([19,15,4]\) structure that contained a \([12,10,2]\) structure than non-targets. Here we found a highly significant difference: 94% (34 of 36) of Staufen target transcripts with a \([19,15,4]\) structure had one containing a \([12,10,2]\) structure but only 61% (174 of 283) of such non-target transcripts did \((P < 2 \times 10^{-5}, \text{Hypergeometric test})\). We also noted a similar, but not statistically significant, preference for \([19,15,0]\) structures containing a \([12,10,0]\) structure in the Staufen targets: 91.7% (22 out of 24) versus 69% (29 of 42). These results indicate that \([12,10]\) structures are not only independent predictors of Staufen binding – although the statistical significance of this association is weak – but also features of \([19,15]\) structures are more likely to be found in Staufen targets.

Based on these results, we define three classes of Staufen recognized structures (SRSs): Type I is a \([19,15,0]\) containing a \([12,10,0]\), which we abbreviate as \([19,15,0]\&[12,10,0]\); Type II is a \([19,15,4]\) containing a \([12,10,2]\), abbreviated \([19,15,4]\&[12,10,2]\); and Type III is a \([12,10,2]\). An example of a Type II SRS is shown in Figure 3.10D.
Figure 3.16. Mapping of dsRNA structures in the 3'UTRs of Staufen targets and non-targets.

The three refined dsRNA stem structures were mapped in the 3'UTRs of Staufen targets (A), length-matched non-targets (B), and a random subset of non-targets (C). The x-axis represents the 3'UTR in nucleotides, starting from the first nucleotide after the stop codon. Each 3'UTR is represented by a grey bar, within which the predicted SRS hits are represented by colored bars ([19, 15, 0]: dark blue; [19, 15, 4]: light blue; [12, 10, 2]: red; [12, 10, 2] embedded in or overlapping with [19, 15, 0]: black; [12, 10, 2] embedded in or overlapping with [19, 15, 4]: magenta). For each SRS, the 5'-most nucleotide in the corresponding 15 of 19 or 10 of 12 motif is connected to the paired nucleotide in the partner arm by a line of the same color as the SRS ([19, 15, 0]: dark blue, [19, 15, 4]: light blue, [12, 10, 2]: red).
Figure 3.17 individually compares the true and false positive rates for each of the structures mentioned above. If we classify transcripts based on which type of SRS they contain, we achieve an AUROC of 81.6% by training with the whole dataset, which is comparable to that achieved on held-out data by a logistic regression classifier that uses presence or absence of all the structures mentioned in the previous paragraph plus 3’UTR length as features. Note that because we used the training set to define these structures and the SRSs, these AUROCs likely overestimate the classification performance we should expect on new data; however, we suspect that this over-estimate is small because i) we use a very simple classification model, ii) the enrichment $P$ values we used to define these features were all highly significant, and iii) we only made a small number of feature selection choices. We also note that $[12,10,0]$ lies directly on the curve, so it has no added classification power in our training set above the three types of SRSs and, as such, we do not identify it as an additional SRS. $[19,15,0]$ lies slightly to the left of the ROC curve formed by the three SRSs, so it does have added classification power; however, the increase is small and, as such, we omit it as a separate feature in order to simplify the model and to ensure that our SRSs are nested (i.e., all Type I SRSs contain Type II and Type III SRSs) thereby supporting our simple transcript scoring scheme.

The crystal structures of dsRBDs from *Drosophila* Staufen or yeast Rnt1p endonuclease bound to dsRNA show interactions between the RBD and the hairpin loop. We, therefore, next asked whether there was any preference for the location of the three types of SRSs relative to hairpin loops. We scored transcripts based on the linear distance on the mRNA transcript between the two paired regions in the SRS, and found that this minimum distance was significantly larger in positive transcripts for Type I SRSs but shorter for Type II and Type III SRSs (Figure 3.18). These results were, therefore, inconclusive with respect to the preferred location of SRSs relative to hairpin loops. However, we note that the median distance for Type III SRSs is particularly small (7 nucleotides), a feature that we consider in the Discussion.
Figure 3.17. Classification of transcripts based on SRS type.

ROC curve to assess how well the SRSs distinguish Staufen targets from non-targets (see Materials and Methods for detailed information). The figure also shows the true- and false-positive rate using the presence and absence of other structural elements, including [19,15,0], [19,15,4], [12,10,0], [12,12,0] and [19,19,0]
Figure 3.18. Distance between the arms of dsRNA stems.
3.3.6  Mapping of SRSs in *Drosophila* and human mRNAs

We next mapped the locations of the three types of SRSs within the 3’UTRs of all *Drosophila* mRNAs (Supplementary Table S3.7). Figure 3.19 shows the locations of SRSs in the 3’UTRs of the Staufen target transcripts; length-matched non-target transcript 3’UTRs; and a random subset of non-target 3’UTRs.

We then asked whether the identified *Drosophila* SRSs predict human Staufen targets better than 3’UTR length alone and found that none had better AUROC than 3’UTR length for either Staufen1 or Staufen2 (Table 3.4). These results, as well as the failure to predict a motif for *Drosophila* Staufen from the GFP two-fold data set, may be a consequence of these lists containing a significant fraction of weakly associated targets or co-expressed transcripts that are not normally bound by Staufen but are scored as bound in these data sets due to overexpression of the tagged Staufen protein used in the RIP-Chip experiments. As such those transcripts may not be enriched for SRSs (see Discussion).
Figure 3.19. Mapping of SRSs in the 3'UTRs of Staufen targets and non-targets.

The three types of SRSs were mapped in the 3'UTRs of Staufen targets (A), length-matched non-targets (B), and a random subset of non-targets (C). The x-axis represents the 3'UTR in nucleotides, starting from the first nucleotide after the stop codon. Each 3'UTR is represented by a grey bar, within which the predicted SRS hits are represented by colored bars (Type I: dark blue; Type II: light blue; Type III: red; Type III embedded in or overlapping with Type I: black; Type III embedded in or overlapping with Type II: magenta). For each SRS, the 5'-most nucleotide in the corresponding 15 of 19 or 10 of 12 motif is connected to the paired nucleotide in the partner arm by a line of the same color as the SRS (Type I: dark blue, Type II: light blue, Type III: red).
<table>
<thead>
<tr>
<th>Datasets</th>
<th>SRSs</th>
<th>3'UTR length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staufen1</td>
<td>85.76%</td>
<td>88.90%</td>
</tr>
<tr>
<td>Staufen2-59</td>
<td>78.52%</td>
<td>87.55%</td>
</tr>
<tr>
<td>Staufen2-62</td>
<td>78.71%</td>
<td>86.14%</td>
</tr>
<tr>
<td>Anti-GFP, 2-fold</td>
<td>68.25%</td>
<td>75.24%</td>
</tr>
</tbody>
</table>

Table 3.4. Comparison of using SRSs and 3'UTR length to predict human Staufen targets and two-fold GFP-Staufen targets.
To assess whether SRSs map to the experimentally determined \textit{in vivo} Staufen-binding sites, we focused on \textit{Drosophila bicoid}\textsuperscript{155,171} and human \textit{ARF1}\textsuperscript{172} (Figure 3.20). In the \textit{bicoid} 3’UTR, the SRSs mapped almost exclusively to the three experimentally determined Staufen-binding regions: for Type II SRSs the precision was 1.0 and for Type III SRSs the precision was 0.94 (baseline precision = 0.47; there were no Type I SRSs). For human \textit{ARF1}’s 3’UTR, again, the SRSs mapped almost exclusively to the two known Staufen1-binding regions: for all three types of SRSs the precision was 1.0 (baseline precision = 0.13). Taken together, these results provide strong evidence that the computationally identified structures correspond to \textit{bona fide, in vivo} Staufen-binding sites.
Figure 3.20. Predicted Staufen-binding motifs map with high precision to the known Staufen-binding regions in *Drosophila bicoid* and human *ARF1* 3'UTRs.

Mapping of the predicted *Drosophila* Staufen-binding motifs to experimentally determined *in vivo* Staufen-binding regions in (A) *Drosophila bicoid* and (B) human *ARF1* 3'UTRs. The yellow shading in the background indicates the regions are important for Staufen binding *in vivo* as defined in those studies. The x-axis presents the relevant region of the 3'UTR in nucleotides, starting from the first nucleotide after the stop codon. The colored lines represent the entire span of each predicted SRS hit that is indicated on the y-axis, mapped onto the 3'UTR sequence. For each SRS, the 5'-most nucleotide in the corresponding 15 of 19 or 10 of 12 motif is connected to the paired nucleotide in the partner arm by a line of the same color as the SRS (Type I: dark blue, Type II: light blue, Type III: red).
3.4 Discussion

Here we have performed RIP-Chip to identify mRNA targets of Staufen in vivo in early *Drosophila* embryos, using a synthetic antibody to immunoprecipitate endogenous Staufen and an anti-GFP antibody to immunoprecipitate transgenic GFP-Staufen. *In silico* analyses of the functions and subcellular localization of these bound transcripts suggested novel roles for Staufen in early embryos. In addition, computational analyses identified double-stranded RNA motifs that are highly specific to Staufen’s binding to its in vivo targets.

3.4.1 A role for 3’UTR length in Staufen-mediated decay?

*Drosophila* Staufen and human Staufen1 and 2 targets have much longer 3’UTRs than co-expressed non-targets. Studies of human Staufen suggest that Staufen-mediated decay (SMD) is dependent on the presence of an upstream termination codon, and therefore Staufen binding sites that direct SMD must be located in the 3’UTR of targets. In addition, in yeast, *C. elegans, Drosophila* and mammals, transcripts with longer 3’UTRs show increased susceptibility to nonsense-mediated decay (NMD), which shares many mechanistic similarities with SMD. It is therefore possible that the longer 3’UTRs of Staufen targets may be important for SMD.

There is not, however, currently any evidence for SMD in *Drosophila* embryos. If SMD were a major role of Staufen then a large fraction of its targets should be degraded, and while 19-25% of the Staufen targets we have identified are degraded in early embryos, this number is very similar to the overall percentage of maternal mRNAs that are cleared.
3.4.2 Structural motifs that predict Staufen binding

Previous \textit{in vitro} studies on the binding of dsRBD3 from \textit{Drosophila} Staufen to artificial RNA substrates\textsuperscript{170}, and \textit{in vivo} studies on the binding of mammalian Staufen to \textit{ARF1} mRNA\textsuperscript{172}, have identified 12 bp and 19 bp Watson-Crick (or G-U) paired stems, respectively, as binding sites for Staufen. In addition, studies of two other dsRBPs have shown that similar 16 bp stems can also act as binding sites for RBPs containing two or more dsRBDs \textit{in vitro}\textsuperscript{149,188}. Our computational analysis of \textit{Drosophila} Staufen’s \textit{in vivo} targets has revealed enrichment for three types of SRSs in their 3’UTRs: Type I $[19,15,0] \& [12,10,0]$, Type II $[19,15,4] \& [12,10,2]$ and Type III $[12,10,2]$. Together the Type I, II and III SRSs map to the previously identified Staufen-binding regions in \textit{Drosophila} \textit{bicoid} and human \textit{ARF1} mRNAs with a precision of $\geq 0.94$.

To our knowledge, this is the first report that uses genome-wide \textit{in vivo} binding data to define motifs – and thus specificity – for a double-stranded RNA-binding protein. Whereas \textit{in vitro} studies have suggested that mismatches in dsRNA stems reduce the degree of dsRBD-RNA binding\textsuperscript{149,170,189}, all three SRSs predicted from our \textit{in vivo} data allow one or more mismatches, and Type II and III SRSs also allow unpaired bases. Consistent with this finding, in \textit{Drosophila} the experimentally identified regions within the 3’UTR of \textit{bicoid} RNA that are required for Staufen binding contain imperfect stems of the length and mismatch number predicted by our motif finding\textsuperscript{155}. Moreover, in other cases where the binding of dsRBDS to endogenous substrates has been studied, bulges or mismatches are often present. For example, mammalian ADAR2 binds a structure in \textit{GluR-2} mRNA referred to as the R/G stem-loop which contains mismatches required for binding\textsuperscript{111,112}. In addition, mammalian PKR binds to a variety of cellular and viral RNAs that contain mismatches\textsuperscript{190-192}. Thus, a perfect Watson-Crick/G-U dsRNA helix is not a prerequisite for dsRBD binding; indeed, we argue that imperfect stems provide the major specificity of Staufen’s binding to its targets. Experimental analyses will be required to test this proposal. We note that our definition of SRSs permits non-canonical base pairs (or the rotation of
unpaired bases) that lead to stable A-form helical stems in the absence of perfect, canonical base pairing. Indeed, the strong preference for mismatches over unpaired bases supports this possibility.

### 3.4.3 How does Staufen recognize and bind stems of different lengths?

In addition to being imperfect stems, the SRSs we have identified are comprised of two major substructures of length 19 and 12, with the latter often contained within the former. One possibility is that the shorter stems mediate the binding of a single dsRBD and the longer stems mediate the binding of two or more dsRBDS, either housed in the same dsRBP or in two simultaneously bound dsRBP molecules. Consistent with the possibility that the longer stems mediate the binding of two or more dsRBDS, three of Staufen’s dsRBDS bind RNA in vitro and structural studies of *Xenopus* Xlrbpa and mammalian ADAR2 have shown that the binding of a single dsRBD involves interactions with only one face of a dsRNA helix, thereby allowing a second dsRBD to bind a different face of the same helix.

A second, not necessarily mutually exclusive, possibility is that that the short Type III SRS represents the binding site for a dsRBD only in the context of a stem-loop structure while the longer SRSs represent binding to stems with no adjacent hairpin loop. With respect to the former, *in vitro* studies of the binding of dsRBD3 of *Drosophila* Staufen to RNA stem-loops have shown that a stem length of 12 bp with a loop is optimal for binding and that, in this context, the dsRBD makes contacts with the loop as well as the stem. Similarly, the structure of the dsRBD from the yeast Rnt1p endonuclease bound to a small nucleolar RNA substrate shows binding to a 13 bp stem capped by a tetraloop, with the protein again making contacts with both stem and loop. With respect to the longer SRSs, the mammalian Staufen1 binding site within ARF1 mRNA appears to be a 19 bp stem with no adjacent hairpin loop (i.e., the
“loop” enclosed by the stem contains multiple hairpin loop structures) and mutations that shorten this stem reduce Staufen binding\textsuperscript{172}. That different dsRBDs in the same protein may have different binding preferences is supported by data from mammalian ADAR2: one of its dsRBDs prefers hairpin loops while the other prefers duplexes that contain mismatches within internal loops\textsuperscript{111}.

### 3.4.4  Staufen levels as a determinant of target mRNA selection

Our data suggest that modest changes in the levels of Staufen might have a significant effect on its compendium of bound mRNAs. Thus, mechanisms that regulate the levels of Staufen could have biologically relevant effects on the processes that Staufen regulates. This has important implications for RIP experiments that utilize over-expression of tagged Staufen proteins, and likely other dsRBPs, in cell lines or \textit{in vivo} as they may lead to identification of spurious target mRNAs. It may be argued that Staufen and other dsRBPs are particularly prone to artefactual target mRNA identification because of their propensity to bind a range of double-stranded motifs rather than to specific sequences. However, given the low complexity and the redundancy within the recognition sites of many sequence-specific RBPs, it is possible that target site selection for many RBPs is strongly influenced by RBP levels.

### 3.5  Materials and Methods
3.5.1  *Drosophila* stocks

*Drosophila* stocks used were: w1118, GFP-Staufen transgenics\(^{158}\) (line: GFP311) and staufen mutants with the genotypes w; stauD3 cn sp / CyO; GFP-Stau2.2FL/TM3 and TM6B, w+; stauD3/CyO\(^{193}\), and P[ry11] staury9 cn1 / CyO; ry506\(^{151}\).

3.5.2  RNA co-immunoprecipitations

For RIPs with synthetic anti-Staufen antibody, synthetic antibodies were expressed and purified as Fabs, and immunoprecipitations were as described\(^{175}\) with only minor modifications. For anti-GFP-Staufen immunoprecipitations for Western blots, RIP-Chip, and qPCR validation experiments, protein G magnetic beads were first blocked\(^{194}\) and immunoprecipitations were then performed using a protocol adapted from Invitrogen’s Dynabeads® Protein G protocol and Roche’s immunoprecipitation protocol for anti-GFP. The RNA retrieved from these extractions was purified and concentrated using the kit and protocol from Zymo Research (CAT#R1015). For anti-GFP immunoprecipitations for comparison of synthetic antibody to anti-GFP RIPs (Figure 3.7C), a slightly different protocol was used, modified to be as similar as possible to the synthetic antibody RIP protocol.

3.5.3  Microarrays

For microarray analysis, double-stranded cDNA was prepared following the protocol described in the NimbleGen Array User’s Guide (Gene Expression Arrays, version 5.0) with minor modifications. For all samples, 500 ng of double-stranded cDNA was labelled with Cy3- or Cy5-tagged random nonamers following the Roche NimbleGen protocol. Labelled cDNA was then hybridized to custom-designed *Drosophila* 4 x 72K NimbleGen arrays (GEO platform number:
Microarray data was analyzed using the Significance Analysis of Microarrays (SAM) function available in the MultiExperiment Viewer (MeV) software application.

### 3.5.4 Data access

The data reported in this study have been deposited in NCBI’s Gene Expression Omnibus and are accessible through GEO Series accession number GSE43418 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43418).

### 3.5.5 Reverse transcription-quantitative PCR

For RT-qPCR, RNA isolated from immunoprecipitates was reverse-transcribed with random hexamer primers and Superscript II reverse transcriptase (Invitrogen). The resulting single-stranded cDNA was subjected to real-time PCR with SYBR green PCR master mix (ABI) using a CFX384 Real-Time System (Bio-Rad). Relative levels of different transcripts were determined using a standard curve.

### 3.5.6 Source of transcript sequences for assessment of UTR and ORF lengths and motif finding

The *Drosophila melanogaster* (BDGP5) and *Homo sapiens* (GRCh37.p6) transcript sequences were downloaded from Ensembl using BioMart in August 2012 (http://www.ensembl.org/biomart/martview/eea40c9db7c1002506d5c766e8772c08). I downloaded all cDNA sequences and defined 3’UTRs as the portion of the cDNA 3’ to the
coding sequence, as defined by Ensembl. When there were multiple isoforms for a gene I used the longest isoform to represent its mature mRNA sequence.

3.5.7 Definitions of secondary structure terms

Throughout this paper, in reference to a specific secondary structure: we use the term “paired” to refer to RNA bases that participate in a canonical base pair (i.e., a Watson-Crick base pair or a G-U wobble); I use the term “mismatch” to indicate two bases that are found across from one another in a secondary structure but are not canonical base pairs; and I use the term “unpaired” to refer to bases that do not have a corresponding partner base on the other strand of a stem. Mismatches only occur in internal loops and unpaired bases occur in either bulge loops or internal loops although the latter need not contain any unpaired bases. For example, the (3,2) internal loop indicated in the schematic in Figure 3.10A contains two mismatches and one unpaired base, whereas the bulge in Figure 3.10A contains three unpaired bases and no mismatches.

3.5.8 Defining $N$ of $M$ motif hits

I estimated the ensemble probability that a region of $M$ bases will contain at least $N$ paired bases using the ensemble probabilities that individual nucleotides within these regions will be paired. First, for the entire 3’UTR, I computed the single nucleotide base-pairing probability using RNAplfold\textsuperscript{76}, with parameter settings $W=200$, $L=150$ and $U=1$ as recommended\textsuperscript{82}. Note that, when folding the 3’ UTR, we also included the 150 transcript nucleotides that comprise the 5’ flanking sequence of the 3’UTR (i.e., the coding region) so that the folding window was not truncated at the 5’ end of the 3’UTR. I then estimated the pairing probability of the $N$ of $M$ motif
using the lowest single-nucleotide probability in the \( M \)-mer, after removing the nucleotides with the lowest \( (M-N)^{th} \) single-nucleotide probability among all the nucleotides except the 5' and 3' closing bases.

When \( M = N \), this estimate is an upper bound on the probability that the entire region is paired; for other values of \( N \), it is a convenient estimate. For each \( N \ of \ M \) pair, which I call a “motif” because it corresponds to a contiguous sequence of bases, I deemed “hits” to be those regions in the top 1% of the \( N \ of \ M \) probabilities across all 3’UTRs. When scoring other regions of the transcript (5’UTR or ORF) I used the 1% cutoffs defined on 3’UTRs to select motif hits.

3.5.9 Discovery of \( N \ of \ M \) motifs that predict Staufen binding

To assess how well the \( N \ of \ M \) motif hits distinguish the Staufen targets from the co-expressed non-targets, I calculated Wilcoxon rank sum \( P \) values and corresponding Area Under the Receiver Operating Characteristic (ROC) curves (AUROCs) for all values of \( N \) and \( M \) where \( M \) ranged from one to 22 and \( N \) ranged from 0.75 \( M \) to \( M \), with a minimum of \( M-N = 4 \). The Wilcoxon rank sum \( P \) values and the AUROCs were computed based on transcript “\( N \ of \ M \) motif hit scores” which, for a given \( N \ and \ M \), were the sum of the probabilities of all \( N \ of \ M \) motif hits in the transcript’s 3’UTR. These Wilcoxon rank sum \( P \) values and AUROCs were compared to those derived from the 3’UTR length as a baseline. Least absolute shrinkage and selection operator (LASSO) sparse logistic regression was performed to further identify the optimal motifs for Staufen binding. The likelihood ratio test was performed to assess the significance of the improved goodness-of-fit of a regression model containing the selected motifs – over 3’UTR length alone – at classifying transcripts according to whether or not they were bound by Staufen. In some instances, I also used these methods to assess the goodness-of-fit of regression models.
based on motif hits in 5’UTRs and coding regions; in those cases, I replaced the 3’UTR-based transcript score with summed probabilities of motif hits in the appropriate region.

3.5.10 Defining \([12,10]\) and \([19,15]\) structures

For each \(10 \text{ of } 12\) and \(15 \text{ of } 19\) motif hit, I input the region and 150 nucleotides of flanking sequence on either side into Sfold\(^{180}\). I used Sfold to compute both the centroid structure of the input sequence as well as 1,000 samples from the structural ensemble. In each structure, I then identified stems that had either a \(10 \text{ of } 12\) or a \(15 \text{ of } 19\) motif hit as one side of the stem and deemed them \([12,10]\) and \([19,15]\) structures, respectively. For a motif hit to be deemed a valid structure, it had to satisfy three criteria: i) at least \(N\) of its \(M\) bases had to be paired, including the first and last bases; ii) its “partner region”, which is the transcript sequence between the bases that pair with the first and last bases of the \(N \text{ of } M\) motif hit, had to pair only with bases in that hit (\(i.e.,\) contain no hairpins); iii) the motif hit had to pair only with bases in its partner region. I found that motif hits that corresponded to valid structures in the centroid also corresponded, in nearly every case, to valid structures in the majority of the ensemble samples. I, therefore, used the centroid to represent the ensemble as this simplified subsequent analysis. I removed any motif hit from consideration that did not correspond to an appropriate valid structure in the centroid. Structures in which both sides of the stem corresponded to motif hits were only represented once in subsequent analyses. These steps are diagrammed in Figure 3.10B and C.

3.5.11 Identification of additional features of Staufen-recognized \([12,10]\) and \([19,15]\) structures

I examined additional features of \([12,10]\) and \([19,15]\) structures to determine whether any of these distinguished the positive and negative sets. They were: i) number of mismatches; ii)
number of unpaired bases; iii) number of bulge loops; iv) number of internal loops; v) maximum loop size – which is the maximum loop size among all bulge and internal loops in the stem spanned by the $[12,10]$ or $[19,15]$ structure where the size of a loop is the length of the longest side of the loop; vi) the distance between the two regions which, depending on the relative position of the motif hit and its partner region, is either the distance between the 3’ end of the motif hit and the 5’ end of its partner region or vice versa. I compared the distribution of these feature values in the positive and negative sets at both the level of individual structures and at the transcript level using cumulative distribution plots and Wilcoxon rank sum tests. The feature value assigned to each transcript was the minimum value for that feature for all corresponding structures in the transcript.

3.5.12 Using Staufen-recognized structures (SRSs) to predict Staufen targets and non-targets

The Results section defines Type I, II and III SRSs; and a Type II SRS is diagrammed in Figure 3.10C. To assess the predictive value of SRSs, we ranked transcripts based on the presence of any of the three types of SRSs in their 3’UTRs. The relative enrichment for transcripts containing the three SRSs is described in the Results: it decreases from Type I to Type III. I therefore assigned the highest rank to transcripts containing a Type I SRS, the second highest to those with a Type II but not a Type I SRS, the next highest to those that only contained a Type III SRS, and the lowest rank to those without any SRSs. I then plotted an ROC curve that demonstrates the ability of this ranking to distinguish positive and negative transcripts and computed the AUROC.
3.5.13 Scoring of the precision of motif mapping

To compute the precision of the mapping of the SRSs, we computed the proportion of nucleotides in motif hits that are in experimentally defined Staufen binding regions in the 3’UTR of interest. “Baseline precision” is the proportion of the region of interest that is in experimentally defined Staufen-binding regions. In the case of Drosophila Staufen’s binding region in the bicoid 3’UTR, the union of the experimentally defined sites was used in the calculation.

3.5.14 Defining bound and unbound sets for the human Staufens

To define the Staufen targets and co-expressed non-targets in human cells, I re-analyzed the published human Staufen RIP-Chip datasets. The total RNA expression data and RIP-Chip data (GSE 8438, GSE 8437) were separately normalized by justRMA package in R, with a custom CDF (http://brainarray.mbnl.med.umich.edu/Brainarray/Database/CustomCDF/15.1.0/ensg.asp). The genes significantly expressed in the sample were determined by the expression level in the total RNA sample (normalized to the mean expression level) using ‘one class analysis’ in the Significance Analysis of Microarrays (SAM) package, with a false discovery rate (FDR) of 5%. The Staufen targets were determined by comparing the expression level in the immunoprecipitated and the control samples, using the ‘two class unpaired analysis’ method in the SAM package (restricted to the significantly expressed genes), with an FDR of 5% and fold-enrichment of five as the cutoff. The co-expressed non-targets were those with significant expression in the ‘one-class analysis’ but which were negatively enriched in the ‘two-class unpaired analysis’ (i.e., enriched in the control RIPs).
3.6 Additional Supporting Data found in the accompanying CD

Supplementary Table S3.1. List of mRNA targets identified by synthetic anti-Staufen RIP-Chip.

Supplementary Table S3.2. RT-qPCR verification of microarray results for selected mRNAs.

Supplementary Table S3.3. List of mRNA targets identified by anti-GFP RIP-Chip.

Supplementary Table S3.4. Length comparisons between the RBP targets and co-expressed non-targets presented in Figure 4 and Supplementary Figure S6.

Supplementary Table S3.5. Cutoff of top 1% of the $N$ of $M$ probabilities across all 3'UTRs.

Supplementary Table S3.6. Details of the double-stranded motif searches in Drosophila 3'UTRs.

Supplementary Tables S3.7. Location of the predicted SRSs in the 3'UTRs of all Drosophila mRNAs.

I mapped the three predicted Drosophila SRSs to the 3'UTRs of all Drosophila mRNAs, and show the mapping results for each motif in separate tables. Numbers indicate the index of the 5'-most nucleotide of a motif hit in the longest transcript of the indicated gene, with the index of the first nucleotide after the stop codon being set to zero. See Excel spreadsheets.
Chapter 4
Development and application of the RNA-READ pipeline

Some sections of this chapter are derived from the following article: Li, X.*, Kazan, H.*, Lipshitz, H.D., Morris, Q., Finding the target sites of RNA-binding proteins. WIRE, submitted. Wei Jiao performed the modification of RNAplfold, under my supervision. I did all the rest analysis in this Chapter. Supplementary Table S4.1-S4.2 can be found in the accompanying CD.
4.1 Introduction

Early fly embryogenesis is largely mediated by post-transcriptional regulation\textsuperscript{199,200}. The basic molecular asymmetries underlying embryonic axis patterning are often controlled by RNA degradation, localization and coupled translational regulation\textsuperscript{199,201,202}. Large-scale experiments have identified groups of transcripts sharing the same post-transcriptional fate: more than one third of maternally deposited transcripts are degraded during the maternal-to-zygotic transition (MZT)\textsuperscript{5}; 71\% of expressed transcripts show specific location patterns in fly early embryos\textsuperscript{6}; and the ribosomal density and occupancy of many transcripts varies during the first half of embryogenesis\textsuperscript{4}. However, few of the trans-acting factors responsible for these events have been identified. Among them, the RBP, Smaug, is a key regulator that is required for elimination of two-thirds of the destabilized maternal transcripts in \textit{Drosophila} early embryos\textsuperscript{5}. Smaug directs destabilization of \textit{Hsp83} mRNA\textsuperscript{203,204} and represses translation of \textit{nanos} mRNA\textsuperscript{7}, both through Smaug recognition elements (SREs).

Understanding how RBPs recognize and bind their target RNAs is the first step towards characterizing the mechanisms by which they conduct post-transcriptional regulation. Unlike DNA, RNA has a complex secondary, and even tertiary, structure that has a major impact on the ability of ssRBPs to access their target sites. In Chapter 2, I showed that computationally calculated target-site accessibility improves the prediction of sequence-specific binding for various RBPs, with a >22\% average relative decrease in error, compared to the traditional approach of using sequence information only\textsuperscript{54}. Furthermore, I found that using this benchmark, I could distinguish among different ways of predicting accessibility, suggesting its use as way of comparing mRNA secondary structure estimates\textsuperscript{103}. In this Chapter, using this benchmark, I observed that (1) replacing RNAplfold-estimated accessibilities with PARS-measured accessibility\textsuperscript{94} does not improve prediction of \textit{in vivo} target selection by RBPs, (2) replacing the accessibility-based scoring system with one that considers the structural context of ssRNA
improves this prediction. I therefore introduced these scoring systems, which consider the structural context of ssRNA, into my motif-discovery model #ATS described in Chapter 2.

In this Chapter, I also present a motif-scanning procedure, which together with the motif-discovery model compose the RNA Regulatory Elements Analysis and Discovery (RNA-READ) pipeline. Over the past few decades, our knowledge of RNA-protein interaction has grown rapidly. Thanks to large-scale experiments, the target sites of many RBPs have been characterized by either identifying the in vivo targets associated with the RBP\textsuperscript{34,36} or by identifying the in vitro binding site of the RBP\textsuperscript{16,30}. A key ongoing challenge is how to take full advantage of this information. Here, I developed a regression-based motif-scanning procedure which searches the entire collection of RNAcompete-defined motifs\textsuperscript{30} to identify consensus sequences, whose involvement in the regression significantly improves the fitting to the data than ones based on the control features alone. I applied this motif-scanner to Drosophila post-transcriptional regulatory categories, and observed dozens of significant associations between the presence of motif matches and specific regulatory outcomes.

4.2 Results

4.2.1 Comparison of experimental and computational methods for prediction of mRNA structure

Kertesz et al.\textsuperscript{94} introduced PARS as an empirical strategy for genome-wide assessment of mRNA secondary structure. This approach complements computational prediction of mRNA secondary structure, which has been shown to help identify binding sites of both microRNAs and mRNA-
binding proteins (RBPs) as well as to improve genome-wide prediction of in vivo binding\textsuperscript{54,61,64,127}.

To select a method for calculation of mRNA secondary structure in RNA-READ, I first compared the ability of PARS\textsuperscript{94} and RNAplfold\textsuperscript{76} (a standard software package widely used to estimate mRNA secondary structure, used in Chapter 2 and Chapter 3) to recover functional binding sites. To assess the relative accuracy of experimentally assayed versus computationally determined mRNA secondary structure, I applied a slightly modified version of my benchmark\textsuperscript{54} to compare how well each set of structures supported the prediction of in vivo RBP binding. Specifically, I compared the ability of PARS and RNAplfold to recover RBP binding sites. RIP-Chip data was used to define sets of bound transcripts (i.e., positives) and co-expressed but unbound transcripts (i.e., negatives) for nine yeast RBPs with defined consensus single-stranded binding motifs that are predictive of in vivo binding\textsuperscript{54}. For each RBP, I scored every bound or unbound transcript according to the structural accessibility of all sites in that mRNA that matched the RBP’s previously described consensus motif (as described in\textsuperscript{54}). Briefly, for each RBP, the accessibility score for an entire transcript was set to be the maximum of the RNAplfold-predicted accessibility scores for each match to the RBP consensus motif in that transcript. The accessibility score of a match was set to be the minimum of the accessibilities of all nucleotides in the match. For many sites, PARS scores were unavailable for every nucleotides, so this minimum was calculated over all nucleotides for which PARS data were available. For PARS, I used the inverse of the PARS score as a measure of single-nucleotide accessibility and, for RNAplfold, I used the predicted probability that the nucleotide was single-stranded. I have previously reported that the minimum is the best single-nucleotide predictor of the accessibility of the entire binding site and the maximum is nearly as good as the sum at consolidating estimates from multiple sites\textsuperscript{54} (see Chapter 2). This slight modification of my original methodology has the advantage of not requiring me to select an arbitrary non-linear transformation of either PARS or RNAplfold scores. For each RBP, I then ranked transcripts
according to their accessibility scores and evaluated how well that ranking distinguished positive from negative transcripts using the Area Under the Receiver Operating Characteristic (AUROC), a standard metric commonly used for this purpose.

Using these methods, I found that RNApIfold-based calculations of site accessibility are significantly better predictors of in vivo binding than those provided by PARS ($P = 0.004$, two-tailed sign-test; Figure 4.1A). Note that only 58% of nucleotides have a defined PARS score, possibly due to the non-uniform ability of V1 and S1 nucleases to cleave different parts of an mRNA and/or insufficient sequencing depth. RNApIfold remains a better predictor than PARS even when only those nucleotides with PARS scores are considered ($P = 0.04$, two-tailed sign-test; Figure 4.1B).

One factor that contributes to the unexpectedly poor performance of PARS is that in vitro refolding may not capture the in vivo structure\textsuperscript{205}. Sequencing errors are likely to lead to additional inaccuracies; reliable quantification of structure profiles may require a higher read count than the one read/nucleotide that was used to define the PARS scores. Indeed, when I restricted my analysis to transcripts with an average of five reads/nucleotide or higher, the performance difference between PARS and RNApIfold was no longer statistically significant ($P = 0.18$, two-tailed sign-test; Figure 4.1C). Unfortunately, very few transcripts have any PARS data at this restrictive threshold.

In summary, whereas PARS provides a useful empirical tool to assess mRNA secondary structure on a genome-wide basis, at present data can be collected for only a subset of nucleotides and coverage is strongly biased towards highly expressed transcripts. On the other hand, computational methods such as those based on RNApIfold provide information on every nucleotide and are not sensitive to transcript abundance.
Figure 4.1. Comparison of prediction accuracy for in vivo binding of nine yeast RBPs using PARS and RNAplfold to estimate the secondary structure of bound versus unbound transcripts.

The results using PARS are shown on the y-axis, those using RNAplfold on the x-axis. (A) The analysis was performed on all consensus sites containing at least one nucleotide with a nonzero PARS score. (B) The analysis was performed only considering nucleotides with non-zero PARS score. (C) As for (B) but with the additional constraint that the transcript load (i.e., reads/nucleotide) was at least five. *P*-values were calculated using the two-tailed Sign-test.
4.2.2 Consideration of specific structural context further improves mRNA target prediction

As shown, using RNAplfold-predicted target site accessibility improves the prediction of in vivo binding for the majority of sequence-specify RBPs\textsuperscript{54}. As such, I asked whether I could further improve this prediction by introducing more detailed structural features of the accessible target sites (\textit{i.e.}, location in a hairpin loop, interior loop, multiloop or exterior loop, see Figure 4.2 for the structures). To do so, Wei and I modified RNAplfold to allow it to output the probability of a site being accessible, as well as the probabilities of this site being embedded in any of these four structures. I applied my benchmark\textsuperscript{54} to compare how well each of the following scoring systems support the prediction of in vivo RBP binding: (1) number of motif matches (mm), (2) expected number of accessible motif matches (amm), (3) expected number of amm in exterior loops, (4) expected number of amm in hairpin loops, (5) expected number of amm in interior loops, (6) expected number of amm in multiloops. Scoring systems (5) and (6) never give better predictive accuracy than scoring systems that used the expected number of accessible motif matches, so only the results from scoring systems that used the expected number of accessible motif matches, so only the results from scoring systems (1), (2), (3) and (4) are shown in Figure 4.2. For six out of nine yeast RBPs, inclusion of the structural context of the target sites improved prediction of in vivo RBP targets. Among them, the RBP, Vts1, prefers motif matches in hairpin loops to motif matches in exterior loops, while the RBPs, Puf3 and Puf4, prefer motif matches in exterior loops. These results are consistent with previous knowledge from crystal structures\textsuperscript{22,110,132}. 
Figure 4.2. Structural context of target sites improves prediction of target mRNAs bound in vivo by RBPs.

Bar graphs compare the accuracy of different methods that use the structural context of motif matches to predict in vivo binding of RBPs. The inset describes the different bars within the graph.
4.2.3 RNA Regulatory Elements Analysis and Discovery (RNA-READ)

As shown in Figure 4.3, I developed the RNA-READ pipeline, which applies a discriminative framework to identify RNA cis-regulatory elements, either by analyzing enrichment of previously defined motifs or through discovery of novel motif consensuses, which distinguish the positive (e.g., regulated transcripts) and negative transcripts (e.g., the co-expressed transcripts not affected by the same post-transcriptional event).

Figure 4.3. Flowchart of RNA-READ pipeline.
4.2.3.1 Motif discovery in RNA-READ

Motif discovery in RNA-READ was developed based on the #ATS model, which is a discriminative motif-finding procedure that attempts to identify consensus sequences that, when scored with #ATS, best distinguish positive from negative transcripts by having more accessible sites among the positive transcripts. RNA-READ adopts the same discriminative framework and the greedy search algorithm applied in #ATS model, but it extends the #ATS model by recruiting four more scoring systems: it assesses not only the expected number of accessible motif matches (amm), but also the expected number of amm in hairpin loops, the expected number of amm in exterior loops, the expected number of amm in multiple loops and the expected number of amm in interior loops. RNA-READ first selects the five hexamers that, when scored by one of the five scoring systems, best distinguish the positive and negative transcripts. RNA-READ then refines the selected hexamer/score system pairs using the same greedy search procedure applied in #ATS model.

I applied RNA-READ to two genome-wide datasets, which identified the set of transcripts associated with Smaug (unpublished data from the Lipshitz lab) and the set of transcripts whose destabilization in early embryo is dependent on Smaug, respectively. To avoid over-fitting, I used a five-fold cross-validation scheme, in which I split the data into five parts of equal size, trained a motif from each of the possible combination of four sets, and tested the predictive power of this motif on the held-out set. As shown in Figure 4.4, RNA-READ successfully recovered the SREs from both data sets.
Figure 4.4. Smaug recognition elements (SREs) are enriched in *Smaug* associated targets and *Smaug* regulated targets in *Drosophila*.

(A) RNA-READ motif discovery results for the *Smaug* RIP-Chip dataset. (B) RNA-READ motif discovery results for the *Smaug*-dependent degradation dataset. The shown motif was obtained by training the whole dataset, and the shown AUROCs are the average AUROCs calculated on the held-out data. The dashed lines in the stem indicate that the stem length is arbitrary.
4.2.3.2 Motif scanning in RNA-READ

A group of mRNAs sharing the same post-transcriptional fate is often controlled by multiple trans-factors; as such, its fate is unlikely to be well-predicted by a single cis-element. My algorithm described above identifies a single cis-regulatory element by separately refining the top-ranked five hexamers that best distinguish the positive and negative sets. However, the refinement process may lead to a degenerate motif when multiple cis-elements are present, as shown in Figure 4.5. I could address this problem by developing a motif finder allowing multiple-element detection or, alternatively, by performing enrichment tests using previously defined cis-elements through a motif-scanning procedure.
Figure 4.5. Preliminary RNA-READ motif discovery results on Fly-FISH data.

Each dot represents a motif discovery result on a localization category (the positive gene list) versus the co-expressed but ubiquitous category (the negative gene list). Motif discovery was run on the training set only. Motifs presented were randomly picked and follow IUPAC nucleotide codes (R: C/G, Y: C/U, S: G/C, W: A/U, K: G/U, M: A/C, B: C/G/U, D: A/G/U, H: A/C/U, V: A/C/G, N: any base). Red in a motif indicates the existence of non-degenerate nucleotides.
The use of defined elements is easier than the development of a motif finder, but is restricted by
the relatively small set of previously defined motifs. RNAcompete is a large-scale experimental
strategy for identifying RBP binding sites in vitro\textsuperscript{16,30}. To date, binding sites for 205 RBPs from
24 species have been identified directly using this method\textsuperscript{30}. Moreover, it has been suggested that
RBDs with more than 70% amino-acid similarity generally have very similar binding sites\textsuperscript{30};
thus, the set of RBP-binding sites can be expanded by inferring the binding site of untested RBPs
using the tested ones. Both the measured and inferred binding sites have been collected in an
online database, cisBP-RNA (catalogue of inferred sequence binding preferences for RNA)
(http://cisbp-rna.ccbr.utoronto.ca/).

I developed a discriminative motif-scanning framework to search for any previously defined
RNA cis-motif that could distinguish transcripts in the positive and negative sets. Compared to
my motif discovery algorithm based on a greedy search, the motif-scanning procedure does not
select the motif through a compete-and-refine strategy. The predictive power of a motif in
distinguishing the positive from the negative set could, for example, derive from differences in
dinucleotide composition or length between the transcripts in the two sets. I therefore formed a
set of control features based on the dinucleotide composition of the given motif and the length of
the target sequence. I compared the goodness-of-fit of two logistic regression classifiers using a
likelihood ratio test: the baseline classifier that used only “control features” and the motif-based
classifier that adds a feature equal to the expected number of accessible copies of the motif in the
transcript region under consideration. To control for over-fitting, I used an initial LASSO logistic
regression to select features used in each classifier. I only reported the motifs that significantly
improved the prediction.

I used previously published \textit{Drosophila} post-transcriptional regulation (PTR) datasets (\textit{i.e.}, the
FlyFISH website\textsuperscript{6} and supplementary data from a variety of published studies\textsuperscript{4,187,199,206}) to define
a set of 112 categories of post-transcriptional fate. For each category I defined two sets of transcripts: a “positive set” and a “negative set”. The positive set consisted of those transcripts with the post-transcriptional fate described by that category and the negative set consisted of those transcripts that were expressed under the same conditions as the positives but were not annotated as having the given fate (see Materials and Methods for detailed information).

Supplementary Table S4.1 shows all PTR/motif pairs with significant associations (FDR < 0.1, corrected for all pairs) and indicates which region of the transcript was scanned. I found that, in 127 cases, the presence of one of the motifs in either the 3'UTR or coding region predicted either transcript stability, translation, or localization. These associations include previously reported associations between the Pumilio motif and both translational repression and destabilization of transcripts in the embryo\textsuperscript{119,206}, new associations between Orb2 and Aret/Bruno motifs and translational inactivation in early embryos that are consistent with previously suggested roles\textsuperscript{207,208}, a newly predicted association between Rbp1 and Rbp1-like sites in the coding region and localization to pole cells, as well a newly predicted association between Msi sites in the coding region and expression in blastoderm nuclei at Stages 6-7 of embryogenesis.

4.3 Discussion

I have developed the RNA-READ pipeline, which takes as input positive and negative gene lists, and outputs single or multiple RNA cis-regulatory elements that distinguish these sets of transcripts. RNA-READ considers both the primary sequence and secondary structural constrains on binding of RBPs. I identified interesting associations between the presence of specific cis-regulatory elements and the post-transcriptional fates of transcripts by applying the RNA-READ pipeline to the post-transcriptional regulation datasets in fly early embryo.
The RNA-READ pipeline has two working modes: (1) *de novo* motif discovery, and (2) a motif-scanning procedure which analyzes enrichment of previously defined motifs. They were designed to solve different problems. The motif-discovery model attempts to find a single *cis*-regulatory element that best distinguishes the positive from the negative transcripts. It applies a greedy search algorithm to refine the initial motif seed by shortening, lengthening, or introducing degeneracy on a single base at a time. This competing-and-refining process is good for finding the best solution to describe the data, but bad when several solutions are required. It is therefore best used for datasets enriched for only a single major *cis*-element, such as transcripts associated with one RBP or transcripts regulated by one RBP. The caveat is that this RBP may function together with additional RBPs or small RNAs; thus a single motif may not suffice.

The motif-scanning procedure, on the other hand, is less affected by the presence of multiple *cis*-elements, and therefore can be applied to solve regulatory events controlled by multiple *trans*-factors. RNA localization is such an event. Transcripts with the same subcellular localization pattern might be controlled by different mechanisms, including directional transport on cytoskeletal elements, random cytoplasmic diffusion followed by trapping, and generalized degradation combined with local protection\(^{209}\). I designed this procedure such that a motif is reported only when it has significant additional predictive power compared to the control features. This is crucial because I often observed predictive power by using the control features only.

The motif-scanning procedure in RNA-READ has been used to test for enrichment of RBP-binding sites defined by RNAcompete, and can be readily be expanded to include additional, previously reported motifs, including microRNA seed regions, structural elements responsible for dsRBP binding, and other RBP binding sites mined from the literature.

The two modes in RNA-READ could benefit each other to improve the overall classification accuracy (Figure 4.6). Instead of only considering the hexamers, one could also initiate the motif
discovery using the motifs selected by the motif scanner. It could shorten the refining process in the motif discovery and also might provide a better initiation than the hexmers. The current version of the motif discovery has two steps. In the initiation step, the algorithm selects five hexamers that, when scored by one of the five scoring systems, best classify the positive and negative transcripts. In the refinement step, the algorithm separately refines the sequence composition of each hexamer using the scoring system selected in the initiation step, and finally reports the motif with the largest classification accuracy among these five refined motifs. One could include in the initiation step the motifs selected by the motif scanner. The goal of the initiation step is then to find the best scoring system for these motifs and to identify the hexamers with a better classification accuracy than these motifs. To avoid obtaining a degenerate motif, the refinement step in the motif discovery could take as input all the motifs and hexamers from the initiation step but only refine one of them each time by freezing the corresponding scores for the others as a fixed noise to each transcript. The resulting motifs could be further analyzed using the motif-scanning. It is better than scanning only the literature-defined motifs, because it could compensate the incomplete motif collection, and it could refine the tested motifs by optimizing their sequence and the structural context.
Figure 4.6 Future direction of RNA-READ.

The dashed arrows between the ‘REGRESSION’ and ‘MOTIF DISCOVERY’ represent the proposed future work.
4.4 Methods

4.4.1 Quantifying target site accessibility with specific structural restrictions

In RNAplfold, the probability of a site being unpaired is a sum of the probabilities that this site is within a hairpin loop, an exterior loop, an internal loop and a multiple loop. Wei and I modified RNAplfold so that it outputs the probability of the site being unpaired (i.e., accessibility), as well as the probabilities of this site being in one of these four loops. In my experiments, I fixed $W=80$ and $L=40$ and set $U$ to either the width of the consensus sequence or to 1 when calculating single-base accessibility. These parameter settings were previously optimized for predicting siRNA binding\textsuperscript{61,77}. When calculating target site accessibilities for a 3′UTR site, I input the entire transcript into RNAplfold to ensure that the target site accessibility for sites immediately downstream of the stop codon incorporated coding sequence.

4.4.2 PTR categories

For each PTR category, transcripts were classified into two sets: (1) the positive set that is composed of transcripts sharing the same post-transcriptional fates; (2) the negative set that is composed of transcripts co-expressed with the transcripts in the positive set, but do not sharing the post-transcriptional fates. Whenever possible, I used positive and negative thresholds established in the original study describing the data. I defined a negative set in cases where no appropriate one was reported in the original study. To ensure that there was sufficient statistical
power for my analysis, I used only PTR categories that contained more than 30 transcripts in each of the positive and negative sets.

4.4.2.1 RNA localization categories

My RNA localization categories were defined based on the FlyFISH database. I followed FlyFISH’s terms to define the positive sets (e.g., “Apical_exclusion_stage6-7”), and defined the negative set using the transcripts that were classified as unlocalized or ubiquitous at the same stage (e.g., “Ubiquitous_Unlocalized_stages6-7”).

4.4.2.2 RNA stability categories

My RNA stability categories were defined based on studies that profiled mRNA expression level in early-stage fly embryos. For RNA stability in PGCs, I followed the categories used in the original paper to define the positive (“PGC_RNA_decay_3-5vs1-3”) and negative sets (“PGC_RNA_stable_3-5vs1-3”). For the study on regulation mediated by the miR-309 cluster in early fly embryos, the original study defined a list of transcripts that are upregulated in miR-309 mutants by comparing expression profiles in 2-to-3 hour miR-309 mutants and 2-to-3 hour wild-type embryos, but did not provide the expression profiles of the rest of the mRNAs in the wild-type embryos at the same time points. I consulted another study to define the negative set, which includes mRNAs that are stable in wild-type embryos between the 2-to-3 hour and 0-to-1 hour time points.
4.4.2.3 Translational repression categories

My translational repression categories were based on a previous study and used to define the positive sets (e.g., “translation_act_0_2h_fly”) and negative sets (e.g., “translation_inact_0_2h_fly”). More specifically, the positive set includes mRNAs actively translated at 0-2 hr (or 4-6 hr or 8-10 hr) embryos, and the negative set includes the mRNAs that are not translated in the corresponding time window.

4.4.3 de novo motif discovery

RNA-READ applies a discriminative framework to model the relative preference between the positive transcripts and the negative transcripts. It is an extension of my #ATS model. That algorithm attempts to identify a consensus site that best distinguishes the positive and negative transcripts by having more sites in a certain structural context (i.e., located in an accessible region, hairpin loop, exterior loop, multiple loop or interior loop) among the positive transcripts. It is a greedy search algorithm using AUROC as the measurer of prediction to find the optimal choice. The initial step is to find hexamers to seed the motif finding. To do so, I scored each transcript using five scoring systems: (1) expected number of accessible motif matches (amm) (2) expected number of amm in hairpin loop, (3) expected number of amm in exterior loop, (4) expected number of amm in multiple loop, (5) expected number of amm in interior loop. In the scoring system “expected number of amm”, each transcript was scanned using a hexamer to identify the sequence matches of the hexamer in the transcript, the accessibility of which was estimated using RNAplfold and summed to represent the enrichment of this hexamer in this transcript. A score of zero was assigned to transcripts without a motif match to the hexamer. Finally AUROC was calculated to assess whether this enrichment score served as a predictor to classify the transcript into the positive set or the negative set. In the other four scoring systems, the same evaluation procedure was applied except the modified version of RNAplfold (see
Section 4.4.1 for the detailed information) was used to estimate the probability of each motif match being in that specific structural context. I assessed all of the hexamers using the five scoring systems and selected the five pairs of hexamer-with-scoring system that had the largest AUROCs. Each selected hexamer-with-scoring system pair was further refined in an iterative procedure that shortened, lengthened, or introduced degeneracy a single base at a time. At each iteration, the modified motif with the largest AUROC calculated by the selected scoring system was chosen to seed the next iteration, until AUROC failed to increase or the associated Bonferroni-corrected Wilcoxon-Mann-Whitney $P$-value was > 0.05. After the motif finding process converged for all five seeds, the motif with the largest AUROC out of the five was returned as the final motif. I generated 25 training/test set splits using a 5 x 5 cross-validation procedure whereby I randomly split the positive and negative sets into five equally sized bins, trained the motif models on the sequences in four of the bins, and evaluated them on the remaining bin. I repeated this random split five times and finally collected 25 test-set AUROCs.

4.4.4 Motif scanning

I performed a log-likelihood ratio test to assess whether any of the motifs from my collection could better distinguish the positive set from the negative set when provided to a regression algorithm that also had access to a control set of features that consisted of all of the dinucleotides contained within the corresponding motif as well as the length of the target sequence. The construction of these regression models is described below. The comparisons between the motif and the control features were restricted to either the 3’UTR or the coding region of transcripts. I scored each 3’ UTR or coding region using a given motif by summing the accessibility of all the target sites, where a target site was defined as a perfect match to the IUPAC representation of the motif (see Supplementary Table S4.2 for IUPAC motifs used in these analysis) and the accessibility of a target site was defined as the average single-base accessibility of the bases in
the site. A score of zero was assigned to those transcripts whose 3’UTRs or coding regions did not contain a motif match. The single-base accessibility was assessed using RNAlfold with parameter settings \( W=80, L=40 \) and \( U=1 \). Although the analysis was applied to the 3’UTR or the coding region, the entire transcript was input into RNAlfold to ensure correct folding close to the start codon and stop codon. I used the glmnet.R package (version 1.8)\(^{211}\) to apply LASSO-penalized logistic regression on each PTR dataset using the feature sets containing the score calculated for one motif and the relevant control features. In the LASSO regression, the hyperparameter lambda (i.e., the regularization strength) was selected through a five-fold cross-validation procedure, from the lambda sequence computed by glmnet using the default settings of \( n\lambda \) and \( \lambda_{\text{min.ratio}} \). The final value for lambda was the one (from the sequence) with the smallest average generalization error across the five-fold cross-validations. I then used this value of lambda with the ‘glmnet.fit’ object on the entire dataset to compute the weights for the features. The features with non-zero weights were selected as contributing most to the prediction. After the non-zero weight features were defined, I trained two standard logistic regression models: one using all non-zero weight features (including the motif) and one that contained only the non-zero weighted control features. I then assessed whether there was a significant difference in predictive power between these two nested models using a log-likelihood ratio test (as per the procedure recommended in\(^{212}\)). I then used these \( P \) values to compute a false-discovery rate using the Benjamini-Hochberg procedure.
4.5 Additional Supporting Data found in the accompanying CD

Supplementary Table S4.1. Results of RNA-READ on post-transcriptional regulation categories

Supplementary Table S4.2. IUPAC motifs defined by RNAcompete.
Chapter 5

Summary and future directions

Some sections of this chapter are derived from the following article: Li, X.*, Kazan, H.*, Lipshitz, H.D., Morris, Q., Finding the target sites of RNA-binding proteins. WIRE, submitted. I wrote all the content in this Chapter except Section 5.2.1, which was drafted by Hilal Kazan and updated by me afterwards.
5.1 Summary of my work

In this thesis, I have described my work on identifying RNA cis-regulatory element to aid in RBP target selection. RNA-protein interactions differ from DNA-protein interactions because of the central role of RNA secondary structure. Some RBDs recognize their target sites mainly by their shape and geometry and others are sequence-specific but are sensitive to secondary structure context. For RBPs with both classes of RBDs, my in silico analysis successfully distinguished mRNA targets and co-expressed nontargets determined by in vivo RIP-Chip experiment (Chapter 2: RBPs with sequence-specificity; Chapter 3: dsRBP Staufen).

In Chapter 2, I described my work on assessing the impact of target-site accessibility in selecting targets of sequence-specific RBPs. I defined target-site accessibility as the probability that the entire target site is unpaired when considering the ensemble of all possible structures \(^{72,76-79}\). I compiled previously published in vivo RIP-Chip datasets from yeast, fly and human to define the bound and unbound transcripts for more than a dozen RBPs. These RBPs contain a range of RBDs with diverse primary sequence binding preferences and biological functions. As a result, target-site accessibility significantly improves mRNA target selection for majority of the RBPs with previously determined sequence-binding preferences (Figure 5.1). Furthermore, I also used this benchmark to compare the different ways of calculating accessibility. The result suggests that many RBPs require the entire target site to be accessible.
Figure 5.1. Target site accessibility predicts in vivo binding for a diverse range of RBPs.

Comparison of accuracy in predicting bound transcripts based on a given consensus, using either #ATS (i.e., the expected number of accessible target sites, y-axis) or #TS (i.e., the number of target sites, x-axis). Each dot represents the results of an RBP coupled with its previously defined consensus sequence. If there are multiple reported consensus sequences for a protein, the result for each is shown and is distinguished from others by a superscript. Cartoons indicate the species of origin (yeast, fly, or human). RBPs in bold have significantly improved AUROC for #ATS versus #TS ($P < 0.05$, Delong-Delong-Clarke-Pearson test). The RBDs housed in the RBPs (using SMART domains) are summarized in the pie graph. Modified from\textsuperscript{54}.
In Chapter 3, I present my effort on finding binding preference for dsRBP Staufen. By computational analysis of *Drosophila* Staufen’s targets and co-expressed non-targets, both of which were defined using genome-wide RIP-Chip experiments, I was able to define two binding preferences for Staufen: (1) Staufen targets generally have long 3’UTRs; (2) Staufen targets are enriched for specific structural elements in their 3’UTRs.

There were two major concerns for my work: (1) the accuracy of computational tools in predicting mRNA secondary structure. (2) the influence of trans-acting factors. In the other words, whether the structure predicted base solely on the mRNA sequence is informative for predicting binding. The success in both Chapter 2 and 3 support the computational prediction of RNA secondary structure and also suggest that intrinsic mRNA secondary structure plays major role in determining RBP’s binding specificity. The first concern is furhter solved by my observation that computational methods and existing large-scale experimental methods show comparable accuray for predicting RBP binding (Chapter 4).

As such, I developed a discriminative motif-finding method called #ATS (i.e., expected number of Accessible Target Sites) that incorporates accessibility (Chapter 2). #ATS fits a degenerate consensus sequence motif model (e.g., CNGG, where N could be any base). uses a greedy heuristic to build its model: it starts from the five hexamers with the largest predictive power and iteratively refines them (by shortening, lengthening, or introducing degenerate bases) until it can no longer improve the discriminative power of the motif. Applying the #ATS model to several RIP-Chip datasets has successfully recovered the previously identified motifs indicating its ability to identify *in vivo* RBP binding sites.

Finally, I introduced RNA-READ pipeline which applies two working modes to identify cis-regulatory element responsible for RBP binding and function (Chapter 4). In the mode of *de novo* motif discovery, I further extended my #ATS motif discovery model by introducing four more scoring systems that consider the structural context of the ssRNA in the binding site. Using
this motif finder, I identified Smaug recognition elements as being enriched in transcripts associated with Smaug and also in maternal transcripts that are destabilized by Smaug, something that previous de novo motif-finding methods failed to do. In the mode of motif scanning, I applied statistic test to select RNAcompete-defined motifs that significantly better distinguish positive- from negative-set transcripts than the dinucleotide composition of the motif and target length. Using this motif scanner, I identified 127 significant associations between the presence of a motif in either the 3’UTR or coding region of a set of mRNAs, and that set’s stability, translation, or localization in early embryos.

5.2 Future direction

My models do not perfectly reproduce the in vivo binding and functional data, which suggests room for further improvement. The specific future directions for each project have been discussed in the corresponding Chapter. The following sections summarize the general challenges, and point to possible improvements that might be made in several areas.

5.2.1 Gapped motif finders for RBPs

Many RBPs have multiple RNA-binding domains: either repeated copies of the same domain or a mixture of distinct domains. For instance, the human PUM1 protein has eight repeats of the Puf domain where each domain recognizes a single nucleotide. The poly(A) binding protein (PABP) has four RRM domains and each pairwise combination of these domains has a different RNA-binding activity. RBD3 and RBD4 of polyprimidine-tract binding (PTB) protein bind RNA with a fixed orientation relative to each other such that a single RNA cannot be bound simultaneously by these two RBDs unless the two cis elements are separated by a linker
sequence (Figure 5.2A)\textsuperscript{214}. The two dsRBDs in ADAR2 bind to distinct locations in the GluR-2 R/G RNA and both are essential for R/G editing\textsuperscript{111,112} (Figure 5.2B).

The unique modular structure of each RBP is crucial for definition of its mode of target recognition, especially for those RBPs equipped with multiple copies of the same RBD, such as ADAR2\textsuperscript{214} and Staufen. Modeling the modular structure of RBPs in identification of their binding sites is crucial. Some work has, for example, modeled gapped DNA motifs, such as GLAM2\textsuperscript{215}. However, finding gapped RNA motifs is much more difficult than finding gapped DNA motifs since one must take into account primary sequence, secondary and even tertiary, structural elements.
Figure 5.2. Three-dimensional structures of multiple RBDs in complex with RNA.

(A) Solution structure of PTB RBD3 and 4 in complex with CUCUCU RNA (PDB: 2ADC). RBD3 and RBD4 have different binding specificity: RBD3 binds YCUNN and RBD4 binds YCN (Y: pyrimidine; N: any nucleotide). RBD3 and 4 interact extensively, resulting in an antiparallel orientation of their bound RNAs, suggesting that the only way to make these two RBDs bind to a single RNA is to separate their sites by a linker sequence. (B) Solution structure of ADAR2 dsRBD 1 and 2 in complex with GluR-2 R/G RNA (PDB: 2L3J). The dsRBDs recognize their targets by the shape and by the primary sequence in the minor groove. Sequence-specific recognition is achieved through a hydrogen bond to the amino group of G (in the GG mismatch for dsRBD1; in the GC pair for dsRBD2) via a β1-β2 loop and via a hydrophobic contact to adenine H2 (in the AU pair for dsRBD1; in the AC mismatch for dsRBD2) via helix α1. The two dsRBDs bind one face of the RNA and cover about 120 degrees of the turn of the RNA helix.
5.2.2 Methods to predict RNA tertiary structure

Most of RNA structure predictors, as well as the corresponding motif-discovery tools, consider pseudoknot-free RNA secondary structures (i.e., structures that only contain nested base pairs). However, tertiary structures formed by pseudoknots can have biological functions. A number of methods have been developed to predict pseudoknots in secondary structures. One, the Maximum Weighted Matching (MWM) algorithm, which finds the optimal set of base-pairs by finding the highest total edge weights in the folding graph whose vertices, edges and edge weights correspond to the bases, potential base-pairs and the likelihood of a base pair, respectively. Others have presented pseudoknot-specific parameters and introduced an extended dynamic programming algorithm capable of predicting pseudoknots. Furthermore, the MC-Sym pipeline infers RNA tertiary structures based on collection of the arrangements of residues compiled from known 3D structures. Most of the tertiary-structure predictors combine theoretical data and experimental observations but still require manual correction. The analysis of RBP binding could benefit from the prediction of tertiary structure. I expect an improvement of accuracy in classifying Staufen targets and non-targets, if the dsRNA stems bound by Staufen could also be characterized on the level of tertiary structure. Currently, accurately predicting the tertiary-structure for short RNAs is possible. However, tertiary structure prediction for long RNAs remains difficult, and motif finding based on tertiary structures is even more challenging.

5.2.3 Combinatorial interactions among RBPs, miRNAs and mRNAs

Extensive interplay among RBPs or between RBPs and miRNAs has been supported by independent pieces of evidence. For example, combinatorial interactions among RBPs and
mRNAs has been inferred from the fact that pairs of RBPs with overlapping targets are significantly higher than expected by chance. It has also been shown that hundreds of motifs have significant recurrent correlation with regulation following transfection of small RNAs: RBP binding sites, such as U-rich sites (consistent with HuD) and AU-rich elements (AREs), are significantly over-represented in the 3′UTRs of the transcripts whose expression is down-regulated or up-regulated after transfection of small RNAs. Furthermore, miRNA sites are significantly enriched in the human PUM1 and PUM2 targets defined by RIP-Chip experiments, and the PUM-binding motifs (UGUAUA) is enriched in the vicinity of the predicted miRNA sites. Despite all of these observations, the mechanisms of the combinatorial interactions are less well understood. Current small-scale studies have suggested two possibilities: (1) The binding of one factor may block the other due to steric hindrance. For example, the RBP, Dead end 1 (Dnd1), positively regulates its targets by counteracting miRNA-mediated repression. Dnd1 accomplishes this by binding to U-rich regions in the 3′UTR of the target and physically blocking access to a miRNA site. (2) The secondary structure of the target mRNA may induce the cooperative binding of trans-acting factors. For instance, the RBPs, PUM1 and PUM2, induce a conformational change in the 3′UTR of P27 mRNA, thus making a target sequence accessible to a miRNA. It remains to be seen how general these types of phenomena are.

Another challenge is how to computationally model these combinatorial interactions. In Chapter 4, I described RNA-READ framework, which identifies cis-regulatory elements enriched in or depleted for the given positive set versus the negative set. RNA-READ performs likelihood ratio test to assess the significance of the improved goodness-of-fit of a regression model containing the select motif – over the control features alone – at classifying transcripts according to their class labels (positive versus negative). It achieves multiple-element discovery by collecting all the motifs with significant $P$-values in the individual likelihood ratio test, with the assumption that the multiple trans-acting factors are independent for binding to the transcript. One could also
capture the additive and multiplicative relationship among the multiple elements by training this model using the entire collection of motifs, rather than one motif each time. One could also apply nonlinear classification algorithms, such as random forest, to detect other nonlinear relationship among the elements.

5.2.4 Experimental strategies for testing putative RNA cis-regulatory element

Experimental strategies should be applied to verify the computationally identified cis-regulatory elements. An In vivo approach is preferred because the original analysis was performed in vivo.

To examine the importance of accessibility in ssRBP-mRNA interactions (novel accessible elements have been reported in Chapter 2), wild-type and mutated transcripts (with desired predicted accessibilities) will be compared for their binding using immunoprecipitation assay. To simplify the system, three types of transcripts (transcripts from the positive set with high accessibility, transcripts from the positive set with lower accessibility, transcripts from the negative set with low accessibility) that have a single consensus site will be compared. To avoid changing the sequence in the binding site itself, mutations will be introduced into the flanking sequence surrounding the element. The mutations will be designed according to computational predictions to achieve the desired accessibility change of the site without any unanticipated disruption of the structure. I expect less binding if the mutated 3’UTRs have lower accessibility compared to the wild type, and visa versa.

One could apply a similar strategy to examine the relevance of SRSs to Staufen binding (as suggested in Chapter 3). I reported three types of SRSs responsible for Staufen binding (Type I: [19,15,0] & [12,10,0], Type II: [19,15,4] & [12,10,2], Type III: [12,10,2]). I also observed that the relative enrichment for transcripts containing the three SRSs decreases from Type I to Type
III. Both of these observations could be tested using *in vivo* experiments. The focus should be on four types of transcripts (transcripts from the positive set with a single Type I SRS in the 3’UTR, transcripts from the positive set with a single Type II SRS in the 3’UTR, transcripts from the positive set with a single Type III SRS in the 3’UTR, transcripts from the negative set with no SRS in the 3’UTR). One could disrupt the existing SRS, by shortening the length of the stem, or producing more unpaired nucleotides, or removing the Type III SRS embedded in Type I and II SRSs. One could also create SRSs in transcript from the negative set using the opposite strategy. Again, the mutations should be designed based on computational predictions. I expected to detect that the correlation between Staufen binding and the presence of the SRSs. One might expect that the binding affinity of transcript containing the three SRSs will decrease from Type I to Type III.
References


Kiebler, M. A. *et al*. The mammalian staufen protein localizes to the somatodendritic domain of cultured hippocampal neurons: implications for its


166 Kim, Y. K., Furic, L., Desgroseillers, L. & Maquat, L. E. Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. *Cell* **120**, 195-208, doi:S0092867404011572 [pii]

167 Dugré-Brisson, S. *et al.* Interaction of Staufen1 with the 5' end of mRNA facilitates translation of these RNAs. *Nucleic Acids Res* **33**, 4797-4812, doi:33/15/4797 [pii]

10.1083/jcb.201108113 (2012).


