2017

Genomics-Based Studies Identify Cis And Trans Acting Post-Transcriptional Regulators

Shawn W. Foley
University of Pennsylvania, foleys@mail.med.upenn.edu

Genomics-Based Studies Identify Cis And Trans Acting Post-Transcriptional Regulators

Abstract
The identity of every organism is stored in its genetic material. Each gene is transcribed into an intermediate RNA molecule, which undergoes complex processing before translation into a functional protein. RNA processing is controlled by RNA binding proteins (RBPs). Each RBP binds to and regulates the processing, stability, and translation of hundreds to thousands of RNA targets, thereby making these proteins essential for organismal development. RBPs bind to their targets by recognizing both the RNA sequence and secondary structure, which is the interaction between complementary RNA sequences within a single molecule. These interactions can be regulated by changing the chemical makeup of RNA nucleotides via covalent modification, thereby altering the secondary structure and RBP-binding of an RNA molecule. Therefore, the interplay between covalent modifications, secondary structure, and RNA-protein interactions regulates the processing and regulation of each RNA transcript. In this dissertation, I have examined these cis and trans acting post-transcriptional regulators to determine their role in RNA processing.

To do this, we have applied a next generation sequencing technique to globally identify RNA-protein interactions and RNA secondary structure in the nuclei of Arabidopsis seedlings. This work has revealed a strong anti-correlation between RNA structure and protein binding.

We next utilized this same technique to help identify RBPs that regulate root hair cell development. Hair cells are located on the root epidermis and are responsible for the uptake of water and nutrients from the environment. Therefore, increasing hair cell number can increase plant survival. During this work, we identified two RBPs that regulate root hair cell fate, one of which functions in the phosphate starvation response pathway. These findings reveal novel pathways involved in this developmental process.

Finally, we examined the role of covalent modifications in RNA processing. By identifying modifications across the nuclear and cytoplasmic transcriptomes, we found broad populations of modifications corresponding to altered stability. These results illustrate the various regulatory roles held by covalent modifications.

Together, this work has advanced the field of post-transcriptional regulation using the model plant Arabidopsis thaliana, by identifying fundamental features of RNA processing, and has raised many questions for future studies to address.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Cell & Molecular Biology

GENOMICS-BASED STUDIES IDENTIFY *CIS* AND *TRANS* ACTING POST-TRANSCRIPTIONAL REGULATORS

Shawn W. Foley

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Brian D. Gregory, Ph.D
Associate Professor of Biology

Graduate Group Chairperson

_____

Daniel S. Kessler, Ph.D.
Associate Professor of Cell and Developmental Biology

Dissertation Committee
R. Scott Poethig, Ph.D. (Chair)
John H. and Margaret B. Fassitt Professor

Gideon Dreyfuss, Ph.D.
HHMI, Professor of Biochemistry and Biophysics

Kimberly L. Gallagher, Ph.D.
Associate Professor of Biology

Stephen A. Liebhaber, M.D.
Professor of Genetics

GENOMICS-BASED STUDIES IDENTIFY *CIS* AND *TRANS* ACTING POST-TRANSCRIPTIONAL REGULATORS

ACKNOWLEDGMENT

The work performed during my thesis research is the result of collaboration and help from many individuals, and would have been impossible without each one of their contributions. I must first acknowledge and thank my thesis mentor Brian Gregory. Although he tricked me into rotating in his lab, I would not be half the scientist I am today without his guidance. I began my graduate research without any background in genomics or computational techniques, and Brian helped to guide me and direct me throughout my research. He has supported both my good ideas and my terrible ideas, giving me the independence to make my own mistakes and learn from them.

I also need to thank the members of the Gregory lab present when I first joined. The combined help of Ian Silverman, Fan Li, Nate Berkowitz, and Lee Vandivier provided a priceless resource for learning and troubleshooting as I began my research. I need to thank Sager Gosai especially, who worked very closely with me on my first two projects. Sager helped to teach me how to code and analyze next generation sequencing data, and without his help Chapters 2 and 3 of this dissertation would not have been possible.

The Gregory lab has undergone many changes over the years, gaining many new members. Xiang Yu, Lucy Shan, Stephen Anderson, Marianne Kramer, Zachary Anderson, Amelia Solitti, and Bishwas Sharma have all joined or rotated in our lab and have made immense contributions to this work. Troubleshooting with my labmates has helped to make this work far more rigorous, and discussing their research has helped to give me new perspectives on my own.

In addition to my labmates I must also thank my collaborators. Nur Selamoglue, Fevzi Daldal, and Ben Garcia have helped me with all of the mass spectrometry and proteomics work over the last four years. Mark Beilstein, Andrew Nelson, Dorothee Staiger, and Eric Lyons are all experts in plant biology and have helped to make the work in this dissertation relevant to the field at large. Lastly I need to thank Roger Deal and Dongxue Wang for developing and optimizing the INTACT system, and welcoming me to Emory University to help teach me this technique.

I have been lucky enough to have a supportive thesis committee, with members attending almost every talk I have given at Penn to show support and offer guidance. Scott Poethig and Kim Gallagher are true experts in plant genetics and plant root biology, and have offered insights that have elevated this work. Both Kim and her graduate student Ruby O'Lexy have been instrumental in my phosphate starvation research. Stephen Liebhaber is an expert in RNA biology and has always offered a critical eye to my work, giving me perspective on the details that I may have missed. And lastly, I have to thank Gideon Dreyfuss and especially his former research associate Ihab Younis, who have offered support and guidance not just through graduate school, but since my freshman year as an undergraduate. Working in the Dreyfuss lab with Ihab introduced me to research, and helped to shape me as a scientist and cement my decision to pursue my doctorate.

I also need to thank my parents, John and Meryle Foley, for their help and support throughout my entire life. My parents have devoted themselves to their children, with my father working seven days a week for years to send his kids to good schools. My parents and my siblings, Raechel, Jaymi, and Patrick have always been supportive whenever I needed help.

Lastly, and most importantly, I need to thank my fiancée Nicole Clapper. I have been very lucky to have someone so supportive by my side. It is easy to get overwhelmed and bogged down during graduate school, but Nicole always kept me grounded and always knows how to make me laugh. Knowing that I had her and the dogs to go home to made any long days or failed experiments worthwhile. I would not have been able to make it through graduate school without her constant love and support.

ABSTRACT

GENOMICS-BASED STUDIES IDENTIFY *CIS* AND *TRANS* ACTING POST-TRANSCRIPTIONAL REGULATORS

Shawn W. Foley

Brian D. Gregory

The identity of every organism is stored in its genetic material. Each gene is transcribed into an intermediate RNA molecule, which undergoes complex processing before translation into a functional protein. RNA processing is controlled by RNA binding proteins (RBPs). Each RBP binds to and regulates the processing, stability, and translation of hundreds to thousands of RNA targets, thereby making these proteins essential for organismal development. RBPs bind to their targets by recognizing both the RNA sequence and secondary structure, which is the interaction between complementary RNA sequences within a single molecule. These interactions can be regulated by changing the chemical makeup of RNA nucleotides via covalent modification, thereby altering the secondary structure and RBP-binding of an RNA molecule. Therefore, the interplay between covalent modifications, secondary structure, and RNA-protein interactions regulates the processing and regulation of each RNA transcript. In this dissertation, I have examined these *cis* and *trans* acting post-transcriptional regulators to determine their role in RNA processing.

To do this, we have applied a next generation sequencing technique to globally identify RNA-protein interactions and RNA secondary structure in the nuclei of *Arabidopsis* seedlings. This work has revealed a strong anti-correlation between RNA structure and protein binding.

We next utilized this same technique to help identify RBPs that regulate root hair cell development. Hair cells are located on the root epidermis and are responsible for the uptake of water and nutrients from the environment. Therefore, increasing hair cell number can increase plant survival. During this work, we identified two RBPs that regulate root hair cell fate, one of which functions in the phosphate starvation response pathway. These findings reveal novel pathways involved in this developmental process.

Finally, we examined the role of covalent modifications in RNA processing. By identifying modifications across the nuclear and cytoplasmic transcriptomes, we found broad populations of modifications corresponding to altered stability. These results illustrate the various regulatory roles held by covalent modifications.

Together, this work has advanced the field of post-transcriptional regulation using the model plant *Arabidopsis thaliana*, by identifying fundamental features of RNA processing, and has raised many questions for future studies to address.

TABLE OF CONTENTS

ABSTRACT ..... **IV**

LIST OF TABLES ..... **IX**

LIST OF FIGURES ..... **X**

CHAPTER 1: INTRODUCTION TO RNA BINDING PROTEINS, RNA SECONDARY STRUCTURE, AND COVALENT MODIFICATIONS **1**

1.1 INTRODUCTION **1**
1.2 RNA-PROTEIN INTERACTIONS **3**
1.2.1 What are RNA binding proteins? 3
1.2.2 How were RBPs first identified? 5
1.2.3 How were RBPs identified across the proteome? 7
1.3 TECHNIQUES TO STUDY RNA-PROTEIN INTERACTIONS **8**
1.3.1 Identifying RNA-protein interactions in vivo 8
1.3.2 Genome-wide methods for identifying RNA-protein interactions 11
1.4 RNA SECONDARY STRUCTURE **13**
1.4.1 What is RNA secondary structure? 13
1.4.2 How is RNA secondary structure formed? 14
1.5 METHODS FOR PROBING RNA SECONDARY STRUCTURE **15**
1.5.1 Physical methods 15
1.5.2 In silico algorithms 17
1.5.3 Nuclease-based methods 17
1.5.4 Chemical-based methods 18
1.5.5 High-throughput structure probing techniques: Nuclease-based methods 19
1.5.6 High-throughput structure probing techniques: Chemical modifiers 22
1.6 POST-TRANSCRIPTIONAL COVALENT MODIFICATIONS **23**
1.6.1 What are covalent modifications? 23
1.7 TECHNIQUES TO STUDY COVALENT MODIFICATIONS **24**
1.7.1 Biochemical based techniques 24
1.7.2 Transcriptome-wide identification of covalent modifications 26
1.8 OUTLINE OF DISSERTATION **30**

CHAPTER 2: GLOBAL ANALYSIS OF THE RNA-PROTEIN INTERACTION AND RNA SECONDARY STRUCTURE LANDSCAPES OF THE *ARABIDOPSIS* NUCLEUS **32**

2.1 INTRODUCTION **32**
2.2 RESULTS AND DISCUSSION **34**
2.2.1 PIP-seq on purified Arabidopsis seedling nuclei 34
2.2.2 The RNA-protein interaction landscape of the Arabidopsis nucleus 40
2.2.3 Patterns of RNA secondary structure and RBP binding are anti-correlated 44
2.2.4 Distinct RNA secondary structure and RBP binding profiles demarcate alternative splicing and polyadenylation 49
2.2.5 The structural landscape of protein-bound RNA motifs 52

2.2.6 Evidence of post-transcriptional operons in the Arabidopsis nuclear transcriptome 54
2.2.7 CP29A localizes to the Arabidopsis nucleus 57
2.3 CONCLUSION 60

CHAPTER 3: A GLOBAL VIEW OF RNA-PROTEIN INTERACTIONS REVEALS NOVEL ROOT HAIR CELL FATE REGULATORS 63

3.1 INTRODUCTION 63
3.2 RESULTS 66
3.2.1 PIP-seq identifies thousands of cell type-specific protein-bound sites 66
3.2.2 Hair and nonhair cells have distinct RNA-protein interaction and RNA secondary structure profiles in shared mRNAs and lncRNAs 74
3.2.3 SERRATE regulates root hair length and hair cell fate in a microRNA-independent and a microRNA-dependent manner, respectively 78
3.2.4 GRP8 regulates root hair cell fate independently of GRP7 87
3.2.5 GRP8 promotes phosphate starvation stress response 92
3.3 DISCUSSION 97

CHAPTER 4: COVALENT RNA MODIFICATIONS CORRESPOND TO TRANSCRIPT STABILITY 102

4.1 INTRODUCTION 102
4.2 RESULTS 105
4.2.1 mRNA-seq and HAMR analysis 105
4.2.2 The nucleus and cytoplasm have distinct epitranscriptomes 109
4.2.3 Covalent modification site corresponds to transcript abundance and stability 113
4.3 DISCUSSION 116

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS 121

5.1 NOVEL INSIGHTS INTO RNA-PROTEIN INTERACTIONS AND RNA SECONDARY STRUCTURE 121
5.1.1 Determining the regulators of nuclear-specific RNA secondary structure 121
5.1.2 Identifying post-transcriptional operons 124
5.2 FURTHER INVESTIGATING THE ROLE OF RBPS IN ROOT HAIR CELL FATE 124
5.2.1 Utilizing PIP-seq to identify RBP regulators of development and stress response 124
5.2.2 Determining the mechanism of SE-dependent hair cell regulation 125
5.2.3 Utilizing GRP8 in crop development 127
5.3 DEFINING THE ROLE OF THE EPITRANSCRIPTOME IN RNA PROCESSING 128
5.3.1 Determining the mechanism of modification-dependent stability 128
5.4 CONCLUDING REMARKS 129

APPENDIX A: MATERIALS AND METHODS 130

A.1 EXPERIMENTAL MODEL AND SUBJECT DETAILS 130
A.1.1 Plant materials 130
A.2 METHOD DETAILS 131

A.2.1 Cross-linking and INTACT purification 131
A.2.2 Western blotting 131
A.2.3 PIP-seq library preparation 132
A.2.4 Total RNA sequencing library preparation to analyze the ratio of spliced to unspliced mRNAs 133
6.2.5 UV Cross-linking analysis of motifs 134
A.2.6 RNA affinity chromatography 134
A.2.7 MS-ready sample preparation 135
A.2.8 RIP-RT-qPCR 135
A.2.9 Measurement of root hair density and root hair length 136
A.2.10 Measurement of RNA stability 137
A.2.11 Measurement of acid phosphatase activity 137
A.2.12 Measurement of phosphate concentration 138
A.2.13 Measurement of anthocyanin 138
A.2.15 mRNA-seq library preparation 138
A.2.16 GMUCT library preparation 139
A.3 QUANTIFICATION AND STATISTICAL ANALYSIS 139
A.3.1 Experiment specific information 139
A.3.2 Read processing and alignment 139
A.3.3 Estimating unspliced transcripts 140
A.3.4 Identification of PPSs 140
A.3.5 Functional analysis of PPSs 140
A.3.6 lincRNA conservation analysis 141
A.3.7 Calculating the structure score statistic 141
A.3.8 Secondary structure and PPS density at upstream Open Reading Frames (uORFs) 142
A.3.9 PPS profiles across canonical start codons for transcripts localized to specific cellular compartments 142
A.3.10 Structure profile at dsRNase- and ssRNase-identified PPSs 142
A.3.11 Analysis of alternatively spliced exons and introns 143
A.3.12 Analysis of alternative polyadenylation sites 143
A.3.13 Secondary structure at RBP binding sites 144
A.3.14 Motif and co-occurrence analysis 144
A.3.15 Structure score profile analysis of mRNAs 144
A.3.16 PPS profile analysis of mRNAs 145
A.3.17 Mass Spectrometry Analyses 145
A.3.18 Spectral Data Analyses and Protein ID 145
A.3.19 HAMR analysis of mRNA-seq data 146
A.3.20 Calculated modification distribution across a metagene 146
A.3.21 Calculating proportion uncapped 147
A.4 DATA AND SOFTWARE AVAILABILITY 147
A.4.1 Chapter 1 Accession Numbers 147
A.4.2 Chapter 2 Accession Numbers 147
A.4.3 Chapter 1 Genome browser view 147
A.4.4 Chapter 2 Genome browser view 147

APPENDIX B: PROTEIN INTERACTION PROFILE SEUQENCING (PIP-SEQ) PROTOCOL 148

B.1 INTRODUCTION **148**
B.2 REAGENTS AND SOLUTIONS **150**
B.2.1 Crosslinking buffer 150
B.2.2 RIP Buffer 151
B.2.3 16x RNase Stop Buffer 151
B.2.4 1x DNase solution 151
B.2.5 DSN Hybridization Buffer 151
B.2.6 DSN STOP Buffer 152
B.3 BASIC PROTOCOL 1: FORMALDEHYDE CROSSLINKING OF TISSUE **152**
B.3.1 Materials 152
B.3.2 Protocol 153
B.4 BASIC PROTOCOL 2: PROTEIN INTERACTION PROFILING **153**
B.4.1 Materials 154
B.4.2 Protocol 154
B.5 BASIC PROTOCOL 3: STRAND-SPECIFIC HIGH-THROUGHPUT SEQUENCING LIBRARY PREPARATION **157**
B.5.1 Materials 158
B.5.2 Protocol 160
B.6 COMMENTARY **167**
B.6.1 Background Information 167
B.6.2 Critical Parameters and Troubleshooting 168
B.6.3 Anticipated Results 169
B.6.4 Time Considerations 170

REFERNCES **171**

LIST OF TABLES

Chapter 1

Table 1.1: Summary of methods used to probe RNA secondary structure..... 16

Appendix A

Table A.1: Plant lines used in this dissertation. 130

LIST OF FIGURES

Chapter 1

Figure 1.1: RBP identity and binding location regulates RNA processing. 4

Figure 1.2: An overview of techniques used to interrogate *in vivo* RNA-protein interactions. 9

Figure 1.3: An overview of techniques used to interrogate transcriptome-wide RNA-protein interactions..... 12

Figure 1.4: Nuclease-based methods for probing RNA secondary structure..... 20

Figure 1.5: Chemical modifier-based methods for probing RNA secondary structure..... 22

Figure 1.6: Antibody based identification of covalent modifications..... 28

Figure 1.7: CMC-based sequencing identifies sites of pseudoruidinylation..... 29

Figure 1.8: HAMR identifies modifications via *in silico* RNA-seq analysis..... 30

Chapter 2

Figure 2.1: INTACT purified nuclei are free of cytoplasmic, ER, and chloroplastic contamination 35

Figure 2.2: Overview of PIP-seq in *Arabidopsis* nuclei..... 36

Figure 2.3: The PIP-seq libraries passed all three quality control checkpoints during library preparation..... 38

Figure 2.4: PIP-seq is a highly reproducible method 39

Figure 2.5: Characterization of Arabidopsis nuclear PPSs..... 41

Figure 2.6: Characterization of Arabidopsis nuclear PPSs..... 42

Figure 2.7: Patterns of protein occupancy and secondary structure at mRNA start and stop codons 45

Figure 2.8: Secondary structure and protein binding landscapes at protein interaction sites and isolated alternative splicing events. 47

Figure 2.9: Patterns of protein occupancy and secondary structure at pre-mRNA splice sites.... 48

Figure 2.10: Protein occupancy and secondary structure landscapes at alternative splicing and polyadenylation sites..... 50

Figure 2.11: The landscape of protein-bound RNA motifs..... 53

Figure 2.12: Clusters of motifs are present in functionally related genes 56

Figure 2.13: Identification of putative RBPs using synthetic RNA motifs..... 58

Figure 2.14: Identification of Arabidopsis RNA interacting proteins..... 59

Chapter 3

Figure 3.1: PIP-seq was performed on highly pure nuclei 66

Figure 3.2: Nuclear PIP-seq identifies cell type-specific RNA-protein interactions..... 67

Figure 3.3: High quality PIP-seq was performed on highly pure nuclei 69

Figure 3.4: PPSs identified by PIP-seq are highly reproducible 71

Figure 3.5: Cell type-specific PPSs are mostly identified by ssRNase treatment and present in transcripts expressed in both cell types. 72

Figure 3.6: PPSs are primarily present in mRNAs, and enriched in the coding sequence 73

Figure 3.7: RNA secondary structure and RNA-protein interactions are anti-correlated in the nuclei of both cell types 74

Figure 3.8: Hair and nonhair cells have distinct RNA-protein interaction and RNA secondary structure profiles. 77

Figure 3.9: Numerous RBPs are identified via RNA-affinity chromatography 81

Figure 3.10: SERRATE regulates hair cell fate and hair length in a partially microRNA-independent manner 82

Figure 3.11: SE influences root hair cell fate independently of the CAPRICE/WEREWOLF transcription factor network. 84

Figure 3.12: SE-bound GGN motif containing genes regulate root hair cell development 86

Figure 3.13: Enriched protein-bound motifs identified in hair but not nonhair cell PPSs localized to the first 100 nt of annotated mRNA 3' UTRs 88

Figure 3.14: GRP7 and/or GPR8 bind the TG-rich motif *in vivo* 89

Figure 3.15: GRP8 regulates root hair cell fate in a GRP7- and CPC-independent manner 91

Figure 3.16: GRP8 functions in the phosphate starvation response pathway 93

Figure 3.17: GRP8 alleviates phosphate deprivation stress 96

Chapter 4

Figure 4.1: HAMR analysis of highly purified nuclear, cytoplasmic, and whole tissue transcriptomes. 106

Figure 4.2: Covalent modifications are highly reproducible with higher abundance in the nucleus as compared to the cytoplasm of plant cells. 108

Figure 4.3: Covalent modifications are tissue and cellular compartment specific 110

Figure 4.4: Nuclear and cytoplasmic epitranscriptomes have distinct modification localization and makeup 112

Figure 4.5: Modification site corresponds to mRNA abundance and stability 114

Figure 4.6: Working model illustrating the role of modifications in post-transcriptional regulation 120

Chapter 5

Figure 5.1: *Arabidopsis* mRNAs undergo a global protein-dependent restructuring upon nuclear export 123

Appendix B

Figure B.1: Overview of PIP-seq protocol 150

Figure B.2: Size selection gel from strand-specific library preparation 169

Figure B.3: Agilent Bioanalyzer 2100 analysis of RNA from protein interaction profiling 170

CHAPTER 1: INTRODUCTION TO RNA BINDING PROTEINS, RNA SECONDARY STRUCTURE, AND COVALENT MODIFICATIONS

This section refers to work from:

- Foley, S.W.*, Kramer, M.C.*, Gregory. B.D. (2017) A survey of RNA-protein interactions and RNA secondary structure in *Arabidopsis*. *WIRES RNA*. In Revision

- Vandivier, L.E.*, Anderson, S.J.*, Foley, S.W.*, Gregory, B.D. (2016) The conservation and function of RNA secondary structure in plants. *Annual Review Plant Biology*. 67:463-88. PMID: 26865341

- Foley, S.W.*, Vandivier, L.E.*, Kuksa, P., Gregory, B.D. (2015) Transcriptome-wide measurement of plant RNA secondary structure. *Current Opinion in Plant Biology*. 27:36-43. PMID: 26119389
  *Indicates co-first author

1.1 INTRODUCTION

The central dogma of biology states that DNA is transcribed into RNA, which is then translated into protein (Crick, 1970). For decades, researchers have focused on understanding the transcriptional regulators necessary for proper organismal development. These studies have led to the discovery of numerous transcription factors necessary for cellular differentiation events (Bernhardt et al., 2003; Schiefelbein, 2003), environmental responses (Herlihy and de Bruin, 2017; Vishwakarma et al., 2017), as well as proper development and survival (Rux and Wellik, 2016; Scott, 2016). However, transcription regulation alone cannot explain the complex regulatory networks necessary to control the many cell type-specific transcriptomes (Klein et al., 2015; Macosko et al., 2015). More recently, post-transcriptional regulation has become recognized as a major contributor to overall cellular gene expression.

It has been shown that as organism complexity increases, so too does the complexity of the transcriptome (Chen et al., 2014). Every eukaryotic RNA transcript undergoes dozens of highly regulated processing events. For instance, messenger RNAs (mRNAs) undergo 5' capping (Dong et al., 2007), splicing (Buratti and Baralle, 2004; Jin et al., 2011; Liu et al., 1995; Raker et al., 2009; Warf and Berglund, 2010), 3' cleavage and polyadenylation (Klasens et al., 1998; Oikawa et al., 2010), RNA editing and covalent modification (Schwartz et al., 2014a; Zheng et al.,

2013), as well as many other processing events before leaving the nucleus. Nuclear export itself (Grüter et al., 1998), subcellular localization (Bullock et al., 2010; Subramanian et al., 2011), translation (Kozak, 1988; Svitkin et al., 2001; Wen et al., 2008), and degradation (Goodarzi et al., 2012) provide additional regulatory steps that a mature mRNA undergoes. This large number of events allows gene expression to be tightly regulated post-transcriptionally in each cell of an organism. It also allows the RNA products of a single protein-coding gene locus to be processed into dozens to thousands of distinct mature mRNAs (Schmucker et al., 2000; Shang et al., 2017), dramatically increasing the protein-coding capacity of the genome. These complex processing events require a cohort of highly specific protein regulators (Glisovic et al., 2008; Wahl et al., 2009). Therefore, this collection of events that makes up the lifespan of an RNA is regulated by a combination of RNA-protein interactions.

RNA binding proteins (RBPs) are a heterogeneous population of proteins found in all organisms that are defined by their ability to bind RNA (Dreyfuss et al., 1984; Glisovic et al., 2008). RNA transcripts are bound by RBPs throughout their entire lifespan, with RBPs functioning in post-transcriptional processing and regulation of these transcripts (Wahl et al., 2009). The identity of the proteins bound to an RNA molecule, as well as where along the transcript they bind, defines its fate (Licatalosi et al., 2008; Wahl et al., 2009). Therefore, determining the identity and binding sites of RBPs along a transcript can provide information about its regulation. RBP binding is primarily determined by two main factors: the primary RNA sequence of the binding sites and the secondary structure of that site.

As RNA is a single-stranded molecule, it is able to form intramolecular base pairs, and thereby fold into complex secondary and tertiary structures. Once folded, the secondary structure of an RNA molecule mediates its stability and functionality (Foley et al., 2015; Li et al., 2012a, 2012b; Vandivier et al., 2013). For example, transfer RNAs (tRNAs) have a distinct secondary structure that ensures they are aminoacylated with the correct amino acid at their 3’ acceptor stem (Giegé et al., 2012). This is just one of many examples where RNA secondary structure is necessary for the proper functionality of an RNA molecule in biological processes. One of the

primary determinants of RNA folding is the primary sequence of the transcript, as base pairing is limited by the location of potential interacting complementary nucleotides. Other determinants of RNA folding include intracellular abiotic conditions, such as osmolarity (Lambert and Draper, 2007), pH (Draper, 2004, 2008), and temperature (Johnsson et al., 2014; Kortmann and Narberhaus, 2012). However, another determinant of RNA folding is covalent modification of individual nucleotides.

Each RNA molecule is comprised of four different nucleotides: adenine (A), cytosine (C), guanine (G), and uracil (U). However, each of these nucleotides can undergo dozens of different covalent modifications, resulting in >100 known nucleotides present at low concentrations in the cell (Cantara et al., 2011; Limbach et al., 1994; Machnicka et al., 2013). These covalent modifications have first been described in tRNA and ribosomal RNA (rRNA), and are known to influence the secondary structure of these molecules (Arnez and Steitz, 1994; McCloskey and Rozenski, 2005; Sprinzl and Vassilenko, 2005). A majority of known covalent modifications have been found to inhibit RNA base pairing, functioning as “structure busters” (Dominissini et al., 2016; Zhao et al., 2017; Zhou et al., 2016). However, there is also a subset of modification that have been shown to promote secondary structure in the cell (Arnez and Steitz, 1994; Newby and Greenbaum, 2002). Therefore, the role of modifications in RNA folding is determined by the identity of the modification, as well as its location along a transcript. In total, post-transcriptional regulation is an intricately controlled series of events determined by the interplay between RNA-protein interactions, RNA secondary structure, and covalent modifications.

1.2 RNA-PROTEIN INTERACTIONS

1.2.1 What are RNA binding proteins?

From transcription to degradation, RNA molecules are bound by varying cohorts of RBPs. These proteins regulate pre-mRNA splicing (Dreyfuss, 1986; Dreyfuss et al., 1993; Fu and Ares, 2014), 3’ polyadenylation (Darnell et al., 1971; Jelinek et al., 1973; Nakazato et al., 1973; Zheng and Tian, 2014), RNA stability (Garneau et al., 2007; Kiledjian et al., 1997; Rajagopalan et al.,

1998) covalent modification (Alarcón et al., 2015; Fu et al., 2014; Jia et al., 2011), and RNA transport (Izaurralde et al., 1997; Köhler and Hurt, 2007; Lee et al., 1996). RBPs are a diverse class of proteins found in all organisms and are defined as containing one or more RNA binding domains (RBDs). These RBDs can interact with single- or double-stranded RNA (ss- or dsRNA), while RBPs are able to interact with other proteins in the cell through secondary domains.

Gene expression is tightly regulated transcriptionally and post-transcriptionally in both a cell type and developmentally specific manner based on the needs of the cell. It is widely accepted that RBPs are critical for nearly every aspect of the post-transcriptional regulatory steps in eukaryotes. Additionally, RBPs can have distinct regulatory effects depending on where they bind along a transcript. Specifically, the same splicing factor can promote inclusion or exclusion of an exon depending on its binding up- or downstream of that exon (Han et al., 2013). Therefore, the identities of RBPs binding to an RNA transcript, and where they bind along the transcript, will determine its processing (**Figure 1.1**). Thus, misregulation of RBPs directly results in aberrant expression of their target mRNAs, subsequently leading to altered protein levels being produced from these transcripts with sometimes detrimental results (Cooper et al., 2009; Reynolds and Cooke, 2005).

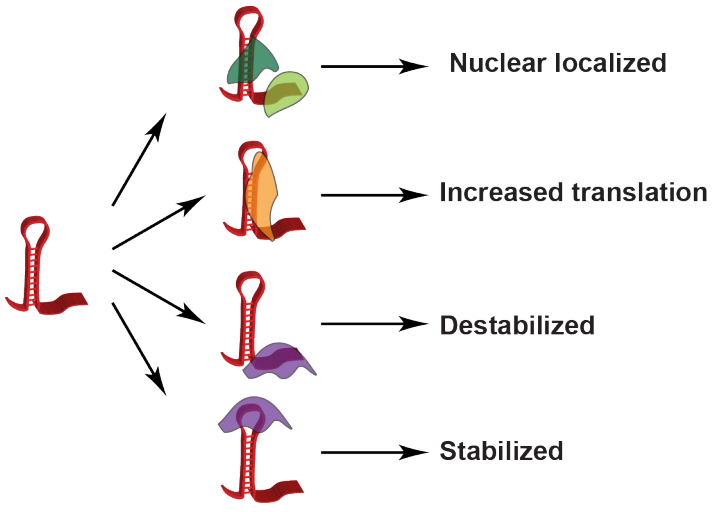


Figure 1.1: RBP identity and binding location regulates RNA processing.
The model illustrates how binding of different RBPs (dark green, light green, orange, or purple) or different binding sites along the transcript can lead to various processing events.

In plants, RBPs have key roles in development (Clarke et al., 1999), response to abiotic stresses (Kim et al., 2005; Kwak et al., 2005), and roles as RNA chaperones (Fedoroff, 2002; Lorković, 2009). As plants are sessile organisms, they must adapt rapidly and efficiently to abiotic stresses to survive. In fact, several known RBPs, such as GRP2 and GRP7, have essential roles in response to cold, both promoting seed germination and conferring freezing tolerance, as well as roles in resistance to salt stress (Kim et al., 2005; Kwak et al., 2005). In fact, plants overexpressing GRP2 germinate substantially better compared to wildtype when exposed to decreased temperatures or increased salt stress (Kim et al., 2007). While RBPs have many implicated roles in many processes in plants, the molecular functions that they have in plant cells are still widely unknown.

1.2.2 How were RBPs first identified?

RBPs were first characterized biochemically by identifying proteins that associated with mRNAs. To do so, researchers crosslinked *in vivo* RNA-protein interactions using 254 nm UV light, and purified polyadenylated RNA by incubating the lysate with a membrane coated with oligo(dT) sequences. The co-purified proteins that bound to the membrane were then eluted and used in two-dimensional polyacrylamide gel electrophoresis (PAGE) that separates the proteins based on both molecular weight and charge. This analysis revealed dozens of different RBPs bound to RNA in nuclei, which were then termed heteronuclear ribonucleoproteins (hnRNPs) (Mayrand et al., 1981; van Eekelen et al., 1981). Additional identification of RBPs was accomplished through RNA affinity purification of RNA-protein complexes coupled with mass spectrometry. With this technique, RNA is first bound to beads and then incubated with a protein lysate. After an incubation to allow any RBPs to bind to the bead-bound RNA, the beads were separated from the lysate and washed to identify any proteins bound specifically to the RNA. Mass spectrometry of these bound proteins was performed to determine the identity of these bound RBPs (Piñol-Roma et al., 1988).

As technologies advanced, RNA immunoprecipitation and cDNA microarrays accompanied with mass spectrometry provided further insights into specific targets of RBPs as well as the composition of protein complexes (McHugh et al., 2014; Tenenbaum et al., 2000). The growing collection of RBP biochemical and structural evidence allowed for the first identification of RBDs through searches for homologous regions in other collections of identified proteins. Some of the most common RBDs identified include the RNA recognition motif (RRM), the K homology (KH) domain, zinc finger motifs, the cold-shock domain, arginine-rich motifs, and dsRNA binding domain (dsRBD) (Glisovic et al., 2008). Interestingly, RBPs may have a single binding domain, multiple domains of the same class, or a mix of several different domains. Completion of sequenced genomes has also allowed further identification of RBPs bioinformatically based on sequence homology to known RBDs (Murzin et al., 1995; Wilson et al., 2009). Through this analysis, the first studies identified ~500 RBPs encoded in the mouse genome (McKee et al., 2005) and ~700 in the human genome (Anantharaman et al., 2002).

RBDs are highly conserved throughout eukaryotic species. While the initial RBD identifications were performed in human, RBPs were identified through RBD homology searches in other species, including *Arabidopsis.* These initial studies in *Arabidopsis* revealed over 200 RBPs encoded in the genome, with the two most commonly occurring RBDs being the RRM and KH domains. Interestingly, over half of annotated RRM containing RBPs in *Arabidopsis* do not have obvious homologs in other metazoans, indicating a possible plant-specific function of these RBPs (Lorković and Barta, 2002). Additionally, there are more than a dozen known RBDs present in *Arabidopsis*, including cold-shock domain, Pumilio, dsRNA binding domains, several types of zinc finger domains, as well as pentatricopeptide repeat (PPR) domains (Lorković and Barta, 2002).

While bioinformatics-based studies using known domains provided the first lists of RBPs, they failed to identify proteins bound to RNA that do not contain a canonical RBD. In order to further identify additional RBPs, researchers have developed techniques that allow for a more proteome-wide identification of RBPs.

1.2.3 How were RBPs identified across the proteome?

Rather than relying on informatics-based approaches, researchers have more recently developed techniques to experimentally identify novel mRNA binding proteins based on direct interaction with RNA, thus providing a more unbiased technique to globally profile RBPs. Through the use of 254 nm UV crosslinking and polyA$^+$ selection, two independent groups have identified several hundred novel RBPs in human cells (Baltz et al., 2012; Castello et al., 2012). Much like the earliest RBP identification assays, this technique uses UV light to crosslink *in vivo* RNA-protein interactions and oligo(dT)-conjugated beads to subsequently purify the polyadenylated RNA population. After isolation, rather than performing a two-dimensional PAGE and identifying proteins individually, this technique utilizes large-scale mass spectrometry analysis to simultaneously probe for all co-purified proteins, thus identifying the RNA binding proteome (Baltz et al., 2012; Castello et al., 2012). This technique can be further enhanced in specificity by using photo-reactive nucleosides, such as 4-thiouridine (4SU) and 6-thioguanosine (6SG). These photoactivatable nucleosides can be readily absorbed by mammalian culture cells and incorporated into nascent RNA molecules (Baltz et al., 2012; Castello et al., 2012). Upon exposure to 365 nm UV light, only these nucleosides become reactive and form covalent bonds with amino acids in the proteins that interact with them. Additionally, when 4SU is exposed to 365 nm UV light, there is a characteristic uridine to cytosine (U to C) transition that can be identified in complementary DNA (cDNA), thus providing evidence of direct RNA-protein interactions as well as evidence of the region of the mRNA bound by the protein.

Both initial studies identified ~800 RBPs in the cell type studied, with ~600 RBPs identified in both studies, indicating the high fidelity of this technique (Baltz et al., 2012; Castello et al., 2012). Both canonical and novel RBPs were identified through these techniques and interestingly, 17 of the non-canonical RBPs identified were enzymes involved in metabolism (Baltz et al., 2012; Castello et al., 2012). Previous studies have suggested that such enzymes may have additional functions outside of their canonical role in metabolism and studies performed

in *C. elegans* (Matia-González et al., 2015) and yeast (Beckmann et al., 2015; Scherrer et al., 2010; Tsvetanova et al., 2010) have also identified numerous metabolic enzymes binding to RNA *in vivo*, indicating they compose a novel class of RBPs.

Recently, this RNA binding proteome capture method was performed in *Arabidopsis* leaf protoplasts and whole leaf tissue. In total, 1,145 mRNA bound proteins were identified, nearly half of which are novel RBPs that do not appear to have canonical or non-canonical RBDs and have not been previously described to bind to RNA. This large number of newly identified proteins for *Arabidopsis* is unsurprising, as many of the identified RBPs are unique to photosynthetic organisms and may be important to chloroplast functions (Marondedze et al., 2016). Furthermore, many RBPs were classified as being involved in response to abiotic stresses, such as response to salt and cold, a classification that matches known functions of several RBPs (Kim et al., 2005; Kwak et al., 2005; Lorković, 2009). Over 200 RBPs were classified as being involved in carbohydrate, energy, and nucleotide metabolism, a theme that appears to be common in RBPs identified in humans, *C. elegans*, yeast, and now *Arabidopsis.*

1.3 TECHNIQUES TO STUDY RNA-PROTEIN INTERACTIONS

1.3.1 Identifying RNA-protein interactions in vivo

Understanding the role of RBPs in post-transcriptional regulation is dependent on identifying their RNA substrates. By understanding which transcripts are bound by an RBP, researchers can determine specific binding sites as well as predict binding preferences based on the interacting RNA sequences. *In vivo* techniques have been developed and applied to better understand true RNA-protein interactions that occur in the cell or organism. RNA immunoprecipitation (RIP) uses an antibody targeting the protein of interest to pull-down the protein as well as any RNAs with which it interacts. RIP can be followed by RT-qPCR to determine if specific RNAs of interest are bound by the RBP or, more recently, RIP has been performed in conjunction with microarray analysis and high-throughput RNA sequencing (RIP-chip and RIP-seq, respectively) (Keene et al., 2006). This provides an unbiased way to identify

RNA molecules bound by a specific RBP (**Figure 1.2A**). This technique has been widely applied in *Arabidopsis,* and has been essential in identifying RNA targets of many plant RBPs (Asakura and Barkan, 2007; Schmitz-Linneweber et al., 2005; Streitner et al., 2012). Despite identifying which transcripts are bound by a protein, RIP does not confer the location along the transcript that the protein is binding. This feature of RIP can be limiting, especially as numerous splicing factors are known to function in a position-dependent manner. Binding of a splicing factor up- or downstream of an intron will promote intron splicing or retention, respectively. Therefore, there was an interest in determining where exactly RBPs of interest may be binding within their target transcripts.

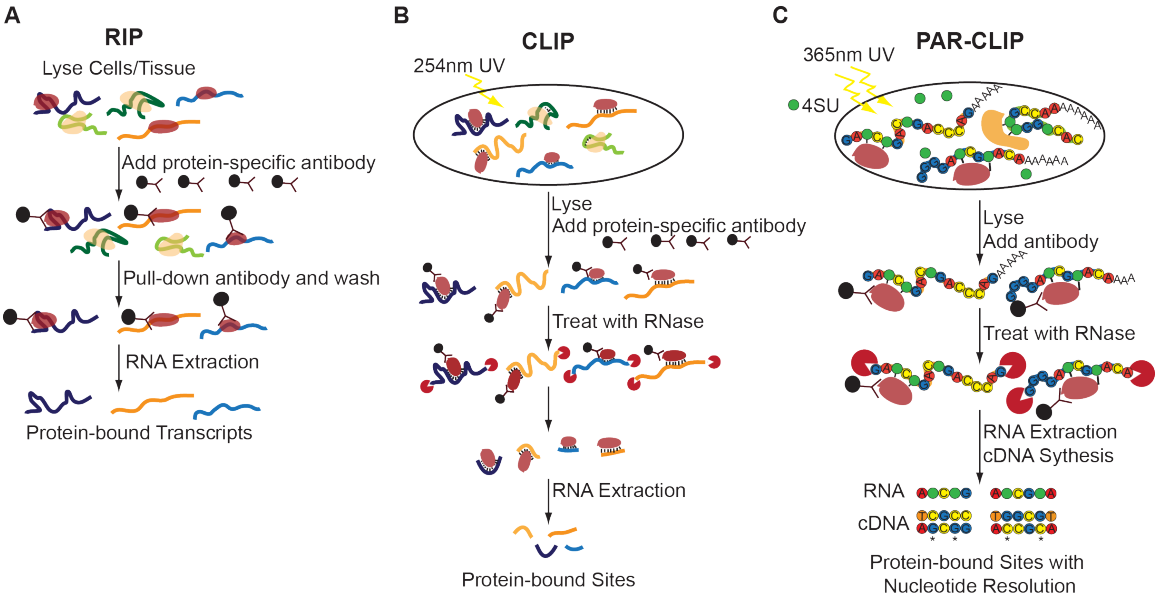


Figure 1.2: An overview of techniques used to interrogate *in vivo* RNA-protein interactions. (A) RNA immunoprecipitation (RIP). RIP can be followed by RT-qPCR (RIP-qPCR), microarray (RIP-chip), or RNA sequencing (RIP-seq). (B) Crosslinking followed by immunoprecipitation (CLIP). CLIP can be followed by RT-qPCR (CLIP-qPCR), or RNA sequencing (CLIP-seq). (C) Photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation (PAR-CLIP).

To that end, crosslinking followed by immunoprecipitation (CLIP) first crosslinks RNA-protein interactions with either UV light or by chemical methods such as formaldehyde. CLIP uses a protein specific antibody to pull down the protein of interest and any bound RNA targets followed by mild ribonuclease (RNase) digestion. These RNases degrade any RNA not bound by

the protein, and thus provides higher resolution view into where along a transcript the protein is binding (**Figure 1.2B**). Similar to RIP, CLIP can be followed by either RT-qPCR or RNA sequencing (CLIP-qPCR or CLIP-seq, respectively) (Licatalosi et al., 2008; Xue et al., 2009; Yeo et al., 2009). The major advantage to CLIP is that it can be used to identify specific protein binding sites. The mild RNase treatment allows for ~30-60 nucleotide resolution of the region of RNA bound by the protein. These regions identified by CLIP-seq can be inputted into motif finding algorithms such as MEME (Bailey et al., 2009) or HOMER (Heinz et al., 2010) to identify RNA binding motifs, which can be used to predict more RNA targets of the gene of interest.

An even higher resolution view of RNA-protein interactions can be achieved via the photoactivatable ribonucleoside enhanced CLIP (PAR-CLIP) approach (Hafner et al., 2010). This method makes use of specialized ribonucleosides that are readily taken up by tissue culture cells and can be incorporated into nascent RNA molecules. When exposed to 365 nm UV light, these nucleosides become reactive and result in the formation of covalent bonds with interacting proteins. This technique has the added advantage that 365 nm UV light only activates these synthetic nucleotides, thereby decreasing false positives due to the crosslinking step. Furthermore, the use of the photoactivatable nucleosides and UV light causes U to C transitions at crosslinked nucleotides, thereby revealing with single nucleotide resolution where proteins are interacting with their target RNAs (**Figure 1.2C**) (Hafner et al., 2010). While PAR-CLIP is an extremely powerful technique to examine RNA-protein interactions, it relies on the ability of cells to uptake the photoactivatable nucleosides and thus can only be performed in cell culture. Because of this, PAR-CLIP has not been performed in *Arabidopsis* or any other plant to date.

Although powerful, immunoprecipitation-based approaches are candidate protein driven techniques and provide only information regarding a single protein of interest. Of late, two genome-wide techniques have been developed to provide a more global picture of RBP-binding across a transcriptome of interest.

1.3.2 Genome-wide methods for identifying RNA-protein interactions

In order to understand global RNA-protein interactions, two techniques have been developed. Global PAR-CLIP (gPAR-CLIP) is a variant technique to PAR-CLIP, which exploits the U to C transitions that occur when 365 nm UV light crosslinks RNA-protein interactions at synthetic nucleotides. Rather than isolating RNA bound by a single protein, this technique identifies any sites throughout the entire transcriptome that exhibit a U to C transition. To do this, cells of interest are incubated with the photoactivatable nucleotide 4SU, which is absorbed by the cell and incorporated into the nascent RNA population of the cells. After exposure to 365 nm UV light, polyadenylated RNAs are isolated via oligo(dT) selection, followed by a light RNase digestion. RNA sequencing allows the identification of RBP-bound regions of RNA on a global scale (**Figure 1.3A**) (Baltz et al., 2012; Freeberg et al., 2013). Using this approach in yeast, over 13,000 protein-binding sites were identified in mRNAs. This analysis further demonstrated that regions bound by RBPs are highly conserved, suggesting the functional significance of the regions of mRNAs that interact with RBPs (Freeberg et al., 2013). The main drawback of this technique is that it relies on the incorporation of photoactivatable nucleosides, and thus cannot be used to study RNA-protein interactions globally in whole tissue samples.

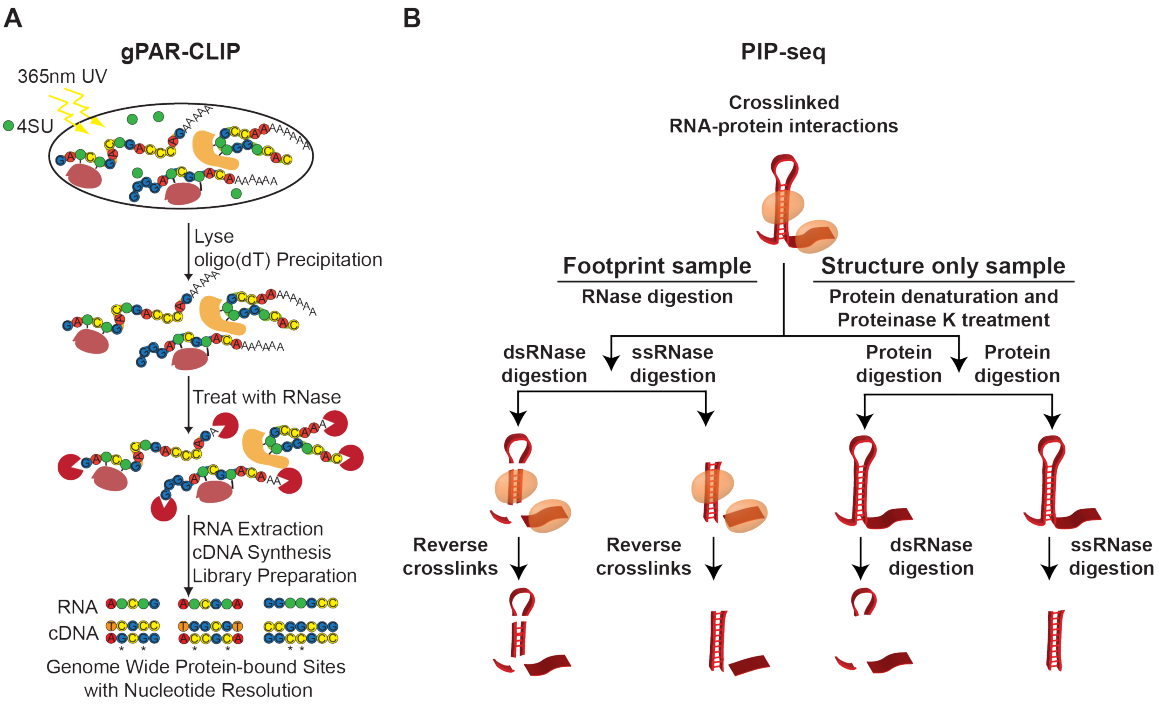


Figure 1.3: An overview of techniques used to interrogate transcriptome-wide RNA-protein interactions.
(A) Global photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation (gPAR-CLIP). (B) Protein interaction profile sequencing (PIP-seq).

An additional technique to globally profile RNA-protein interactions is protein interaction profile sequencing (PIP-seq), which does not rely on incorporation of photoactivatable nucleosides. PIP-seq is able to identify RNA-protein interaction sites within unprocessed and mature RNAs in an unbiased manner. This technique uses formaldehyde as a crosslinking reagent, to circumvent the need for photoactivatable nucleosides. To perform PIP-seq, samples are separated into two sets of sequencing libraries. In the structure only sample all proteins are degraded, and the remaining RNA is treated with a ds- or ssRNA-specific nuclease (dsRNase or ssRNase). This serves as a background for the footprinting sample, in which RNase treatment occurs before protease treatment. Therefore, any regions that is protein-bound in the footprinting sample will not be digested by the RNases and can subsequently be subjected to RNA sequencing. Sequences enriched in the footprinting, with respect to the structure only, samples correspond to protein-bound sites (**Figure 1.3B**) (Gosai et al., 2015; Silverman et al., 2014).

The first PIP-seq study performed in two human cell lines (HeLa and HEK293T) revealed

that the majority of RBP binding in these transcriptomes occurs in introns and coding sequences (CDSs) (Silverman et al., 2014). Similar to what was found with gPAR-CLIP, the protein-bound regions of human RNAs were significantly more conserved compared to adjacent flanking regions. Motif analyses of regions of RNA bound by proteins identified known RNA-bound motifs, such as those that are recognized by RRM domains, as well as novel RBP-interacting motifs. Furthermore, these analyses revealed that protein-bound regions of RNA were enriched in human disease-associated SNPs, indicating a possibly important role of RBPs in the manifestation of human diseases (Silverman et al., 2014). Together, these techniques can reveal fundamental features of RNA-protein interactions across a variety of biological systems.

1.4 RNA SECONDARY STRUCTURE

1.4.1 What is RNA secondary structure?

All RNAs have the capacity to base pair via Watson-Crick, Hoogsteen, or sugar-edge patterns of hydrogen bonds (Leontis and Westhof, 2001; Schroeder et al., 2004). Intermolecular RNA base pairing underlies the coding and replicative abilities of RNA, and enables RNA to serve as a specificity factor in guiding the activity of processes like RNA-directed DNA methylation (RdDM) and microRNA-mediated gene silencing. Intramolecular RNA base pairing is the basis of RNA secondary structure, and is a critical determinant of overall macromolecular folding. In conjunction with cofactors and RBPs, secondary structure forms higher order tertiary structures and confers catalytic, regulatory, and scaffolding functions to RNA. In turn, disrupting the secondary structure of both coding and noncoding RNAs can cause widespread physiological perturbations. For instance, improper tRNA folding disrupts its intricate set of interactions with tRNA synthetases, cofactors, and the ribosome that are required for translation, thus impeding a process fundamental to life (Bhaskaran et al., 2012; Demeshkina et al., 2010). Secondary structure is known to be equally necessary to the functions of ribosomal RNAs (rRNAs) (Nissen et al., 2000; Ramakrishnan, 2014; Steitz and Moore, 2003; Yusupova and Yusupov, 2014), small nuclear RNAs (snRNAs) (Fica et al., 2013; Madhani, 2013), small nucleolar RNAs (snoRNAs)

(Ganot et al., 1997; Kiss, 2002; Kiss-László et al., 1996; Lestrade and Weber, 2006; Ni et al., 1997), and microRNAs (miRNAs) (Carthew and Sontheimer, 2009; Chapman and Carrington, 2007; Kurihara and Watanabe, 2004; Park et al., 2002; Reinhart et al., 2002). Additionally, recent studies are beginning to demonstrate the importance of structure in long noncoding RNAs (lncRNAs) (Novikova et al., 2012; Ponting et al., 2009; Ulitsky et al., 2011; Wang and Chang, 2011) and mRNAs (Ding et al., 2014; Gosai et al., 2015; Li et al., 2012a, 2012b; Rouskin et al., 2014). Thus, a complete understanding of the regulation and functionality of RNAs will require methods to probe and manipulate RNA secondary structure. Here, we review these methods in the context of the form, origins, and function of RNA secondary structure.

1.4.2 How is RNA secondary structure formed?

As with protein folding, the formation of RNA secondary structure is not a simple matter of maximizing the number of stable chemical bonds to minimize free energy. Instead, RNA secondary structure is constrained by transcription, steric crowding, RBPs, covalent modifications and interacting ions. For instance, RNA folding is co-transcriptional, leading to "sequential folding" that can vary with the speed of RNA polymerase elongation (Schroeder et al., 2004). Moreover, RNA folding is guided by proteins and ribozymes with RNA chaperone activity during its initial formation to avoid "kinetic folding traps" (local free energy minima) and improper conformations (Kang et al., 2013; Lorsch, 2002; Mohr et al., 2002; Schroeder et al., 2004; Tompa and Csermely, 2004). Thus, the correct *in vivo* structure of RNA may differ substantially from structures that spontaneously form *in vitro* or the minimum free energy (MFE) structures predicted *in silico*.

Chaperones are a diverse group of proteins functionally defined through their ability to facilitate RNA or protein refolding. RNA refolding is sometimes facilitated by ATP-dependent DEAD-box helicase domains (Lorsch, 2002; Mohr et al., 2002), but can also occur in the absence of external energy. Since chaperones are characterized by their abundance of disordered amino acids, a passive "entropy transfer" model has been proposed in which chaperones adopt the disordered conformation of actively folding RNAs, thus stabilizing RNA folding intermediates and

enabling a more complete “conformational search” (Kang et al., 2013; Schroeder et al., 2004; Tompa and Csermely, 2004). Regardless of their mechanism, chaperones generally lack sequence specificity and possess a wide array of potential targets (Kang et al., 2013; Schroeder et al., 2004; Tompa and Csermely, 2004). As a result, loss of chaperone activity usually causes widespread misfolding and pleiotropic phenotypes (Schroeder et al., 2004).

1.5 METHODS FOR PROBING RNA SECONDARY STRUCTURE

1.5.1 Physical methods

The earliest studies of RNA folding were designed to characterize the three dimensional shape of both prokaryotic (Kim and Rich, 1968) and eukaryotic (Kim et al., 1974; Robertus et al., 1974) tRNAs via X-ray crystallography. The high degree of structure and short length of tRNAs allows them to form crystallized structures more easily than other classes of RNA (Holbrook and Kim, 1997). Although it was a powerful technique in the early studies of RNA secondary structure, X-ray crystallography is limited to transcripts that readily form crystals (**Table 1.1**). Outside of small, highly structured RNAs like tRNAs, there are few classes of RNA that can be readily studied using this approach. Additionally, this technique utilizes *in vitro* folded transcripts, providing only a snapshot of the most energetically stable structure that forms in the buffer tested.

	Method	Biases/Limitations	RNA specificity	Mechanism of Method	Experimental Context
Chemical Modifiers	DMS	A,C specific	ssRNA	Alkalates the N-1 in A and the N-3 in C	*in vivo* and *in vitro*
	Diethyl Pyrocarbonate	A specific	ssRNA	Carboxylates N-7 in A	*in vitro*
	Hydrazine	U specific	ssRNA	Nucleophilic attack of U, removing the base	*in vitro*
	NAI/NAI-N_3	No bias, labels ribose sugar	ssRNA	Acylates 2' hydroxyl of unpaired nucleotides	*in vivo* and *in vitro*
Nucleases	RNase A	Cleaves after purines	ssRNA	Leaves 5'OH and 3'P	*in vitro*
	RNase T1	Preferrential cleavage after guanisines	ssRNA	Leaves 5'OH and 3'P	*in vitro*
	RNase U2	Cleaves after pyrimidines	ssRNA	Leaves 5'OH and 3'P	*in vitro*
	RNase V1	None	dsRNA	Leaves 5'P and 3'OH	*in vitro*
	Nuclease P1	None	ssRNA	Leaves 5'P and 3'OH	*in vitro*
	Nuclease S1	None	ssRNA	Leaves 5'P and 3'OH	*in vitro*
	RNase I	None	ssRNA	Leaves 5'OH and 3'P	*in vitro*
Other	NMR	None	N/A	Aligns molecules in a magnetic field	*in vitro*
	X-Ray Crystallography	Must be crystallizable RNA, in vitro folding only	N/A	Scatters X-rays in an interpretable pattern around an RNA crystal	*in vitro*
	In silico algorithms	Difficult to predict in vivo folding	N/A	mostly predicts based on free energy and conservation	*in silico*

Table 1.1: Summary of methods used to probe RNA secondary structure.
Three categories of methods are covered; chemical modifiers (red), RNases (green), and others (blue).

Conversely, the dynamics of RNA folding can be characterized using solution-state nuclear magnetic resonance (NMR). As opposed to crystallography, NMR can examine the dynamics of RNA folding. Early studies have focused on identifying dynamic secondary structure rearrangement in the lead-dependent ribozyme during autolytic cleavage (Hoogstraten et al., 1998), and in the U6 snRNA (Blad et al., 2005). More recent techniques have allowed greater resolution, allowing characterization of conformational changes on the picosecond time scale (Bothe et al., 2011; Zhao and Zhang, 2015). To date, both NMR and X-ray crystallography are still considered the gold standard in RNA secondary structure probing, revealing the three-dimensional shape of the transcript. However, they are very time and labor-intensive techniques, requiring exhaustive tests in numerous buffer conditions. These limitations prevent such physical methods from being utilized on a large scale.

1.5.2 In silico algorithms

Most algorithms to computationally predict RNA folding patterns are based on minimizing free energy (Gruber et al., 2008; Mathews, 2014; Zuker and Stiegler, 1981). Though widely used, many of these algorithms do not account for protein interactions, evolutionary sequence conservation, or RNA dynamics (**Table 1.1**). Additionally, the fidelity of *in silico* techniques is known to decrease with increasing RNA sequence length, often failing to reproduce known rRNA structures (Zuker and Stiegler, 1981). As opposed to earlier algorithms, the Rfam algorithm offers some improvement by prioritizing the structure of evolutionarily conserved nucleotides, leading to higher fidelity (Griffiths-Jones et al., 2003). However, Rfam is still limited by sequence length, and its database of secondary structure does not include models for any full mRNA molecules. Therefore, experimentally probing structure is necessary to produce reliable models for mRNA folding.

1.5.3 Nuclease-based methods

Early studies of RNases revealed that many of these enzymes specifically cleave ssRNA. This discovery led to nuclease-based footprinting experiments to describe secondary structure of tRNAs (Chang and RajBhandary, 1968). In these experiments, the tRNAs were treated with very low concentrations of a ssRNase in order to induce a single cleavage event within each transcript. This resulted in a population of partially digested transcripts, with each one terminating on a single-stranded nucleotide.

These fragments were then analyzed via end labeling, primer extension, and Sanger sequencing. During an end labeling experiment, the 5' phosphate group is removed from the tRNA via phosphatase treatment, followed by addition of a radiolabeled phosphate through a polynucleotide kinase (PNK) reaction utilizing ^{32}P-γ-ATP. This allows the 5' end of each tRNA fragment to be visualized on film after separation via PAGE. Alternatively, the 3' end of a fragment can be visualized via primer extension. In this technique, a radiolabeled primer is used

in a reverse transcriptase (RT) reaction. The resulting DNA is then labeled near its 3' end, and can be visualized via PAGE. These fragments can then be extracted from the gel and undergo Sanger sequencing (Ehresmann et al., 1987).

In addition to secondary structure biases, RNases can have nucleotide biases, preferentially cleaving after one or more nucleotides. ssRNases with such a bias include RNase A, T1, and U2. RNase T1 preferentially hydrolyzes after guanosines (Loverix and Steyaert, 2001), while RNase A and U2 cleave after purines and pyrimidines, respectively (**Table 1.1**) (Uchida et al., 1970; Volkin and Cohn, 1953). In contrast, nuclease P1, nuclease S1, and RNase I are ssRNases which cleave after each single-stranded nucleotide with equal efficiency (**Table 1.1**) (Desai and Shankar, 2003; Knapp, 1989). These latter enzymes are therefore the preferred ssRNases for footprinting assays.

Although there are numerous ssRNases that can be used in footprinting assays, to date only one dsRNase has been identified and used in such experiments. Isolated from the venom of the *Naja oxiana* (Caspian cobra), RNase V1 preferentially cleaves dsRNA without nucleotide bias (**Table 1.1**) (Favorova et al., 1981; Lockard and Kumar, 1981). The enzyme has been shown to bind double helical RNA before cleavage, so it can also induce cleavage at single-stranded nucleotides within highly structured regions, such as bulges within an RNA stem loop. Overall, when used in conjunction with ssRNases this enzyme has helped to produce a higher resolution image of secondary structure in several tRNAs and rRNAs (Andersen et al., 1984; Favorova et al., 1981; Lockard and Kumar, 1981).

1.5.4 Chemical-based methods

A third method of experimentally probing RNA secondary structure uses chemical modifiers to alter single-stranded nucleotides. One of the first modifiers used was dimethyl sulfate (DMS), which alters unpaired adenines and cytosines. DMS was initially used in conjunction with diethyl pyrocarbonate and hydrazine, which modify adenosine and uridine respectively (**Table 1.1**), to label ssRNA in the yeast 5S rRNA and $tRNA^{Phe}$ (Peattie, 1979; Peattie and Gilbert, 1980).

Aniline was then used to induce strand breakage at the modified bases, allowing mapping via 5' end labeling and PAGE. Subsequent studies revealed that RT cannot process these modified nucleotides, leading to cDNA products terminating at the previous nucleotide, and allowing mapping via primer extension (Inoue and Cech, 1985; Lempereur et al., 1985). Although these early studies were performed *in vitro*, DMS has been shown to easily enter living cells (Lawley and Brookes, 1963), labeling chemical modifier-accessible nucleotides *in vivo* (Antal et al., 2002; Ares and Igel, 1990; Harris et al., 1995; Wells et al., 2000; Zaug and Cech, 1995).

In addition to occluding chemical addition to the nucleoside, base pairing limits accessibility of the 2' hydroxyl group on the ribose sugar (Merino et al., 2005). This limited accessibility is utilized in the selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE), in which 2-methylnicotinic acid imidazolide (NAI) covalently modifies the 2' hydroxyl of the ribose on unpaired nucleotides (**Table 1.1**) (Merino et al., 2005; Wilkinson et al., 2006). Unlike DMS and other nucleoside labeling based techniques, SHAPE labels the ribose sugar, and therefore has no nucleotide bias. Using a single reagent to label each accessible single-stranded nucleotide allows a higher resolution picture of the secondary structure of a transcript than DMS, diethyl pyrocarbonate, or hydrazine alone. However, dsRNA labeling chemicals are not currently available. While ssRNA can be directly identified by these approaches, paired bases are simply inferred by the lack of data from unlabeled nucleotides

1.5.5 High-throughput structure probing techniques: Nuclease-based methods

High-throughput sequencing techniques have revolutionized the study of RNA secondary structure. Several methods have been developed to investigate the structural landscape of eukaryotic transcriptomes. These methods utilize structure-specific nucleases or chemical adducts to identify single- or double-stranded nucleotides (**Figures 1.4 and 1.5**).

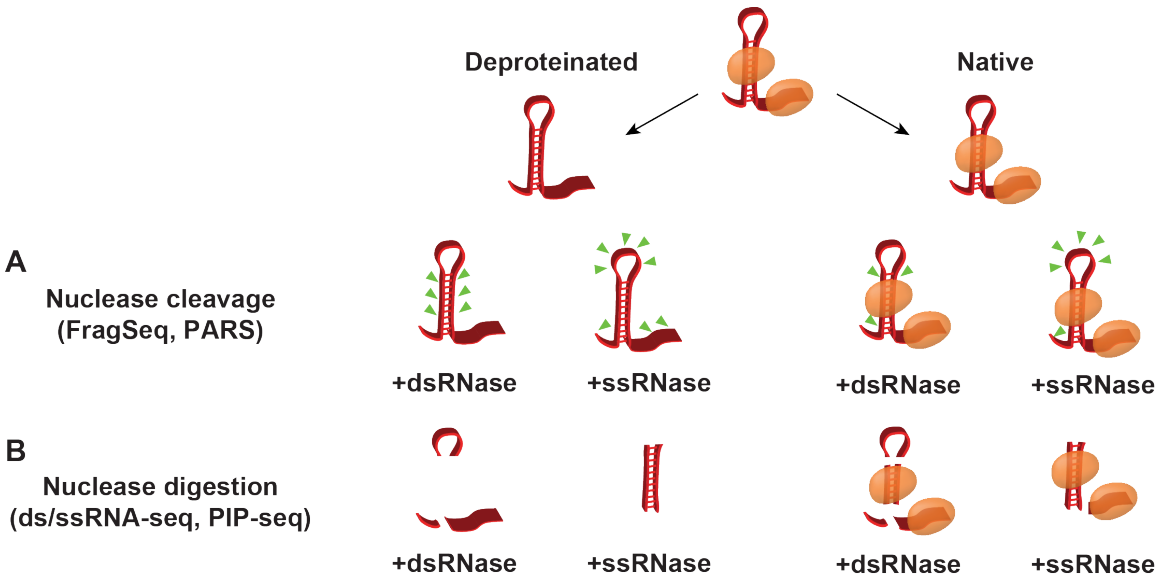


Figure 1.4: Nuclease-based methods for probing RNA secondary structure
A schematic representation of the nuclease-based probing techniques to empirically determine secondary structure. RNA can either be probed in a native state bound by RBPs (orange ovals) or deproteinated through extraction protocols or Proteinase K treatment. (A) PARS assigns structure by the sites of transcript cleavage (green triangles), whereas (B) dsRNase/ssRNase-seq and PIP-seq both work by complete digestion. While multiple cleavage sites are represented in this schematic, it is worth noting that PARS works with single-hit stoichiometry, with one cut interrogated per sequencing read.

Structure-specific RNases have been applied in a transcriptome-wide manner to reveal the global landscape of RNA secondary structure. One of the earlier genome-wide structure probing techniques was FragSeq (Underwood et al., 2010), which utilized nuclease P1 to cleave ssRNA in mouse cells. This cleavage event leaves a 5' phosphate on the RNA fragment (Kuninaka et al., 1961), enabling its selective cloning and sequencing (**Figure 1.4A**). The 5' most nucleotide of these reads therefore corresponds to an unpaired nucleotide, revealing ssRNA across the transcriptome with single nucleotide resolution (Underwood et al., 2010). While powerful, this technique only identifies single-stranded regions, inferring dsRNA from a lack of reads. Other nuclease-based techniques have improved upon this method, identifying both single- and double-stranded regions.

A second nuclease-based approach is the parallel analysis of RNA structure (PARS) technique, which used both ss- and dsRNases to probe structure in yeast (Kertesz et al., 2010) and human tissue culture cells (Wan et al., 2014) (**Figure 1.4A**). To do this, the authors extracted

polyadenylated RNA, which was subsequently denatured and allowed to reanneal *in vitro*. The renatured RNA was then treated with a dsRNase (RNase V1) or ssRNase (nuclease S1) with single-hit stoichiometry, and the resulting fragments underwent high-throughput sequencing to reveal the sites of cleavage. Structure is defined as the ratio of coverage in dsRNA and ssRNA libraries, an estimate of the likelihood for a region to be single- or double-stranded. Unlike FragSeq, this technique provides a single nucleotide resolution view of both single- and double-stranded nucleotides.

The first high-throughput secondary structure analyses in plants were both nuclease-based. These studies utilized the combination of dsRNA-seq and ssRNA-seq, in which RNA is treated with either RNase I (an ssRNase) or RNase V1 (a dsRNase), respectively (**Figure 1.4B**). Unlike PARS, this technique fully digests ss- or dsRNA in a sample to allow every nucleotide of a sequencing read to be informative, offering greater sequencing depth at the expense of resolution. In contrast, PARS and FragSeq only interrogates the structure of a single nucleotide per read. Like PARS, structure is defined by the ratio of dsRNA-seq to ssRNA-seq sequencing depth.

Although informative, each of these initial techniques required the denaturing and reannealing of RNA *in vitro*, thereby interrogating the folded RNA in a protein free environment. PIP-seq is a recently developed technique that identifies RNA secondary structure in its native, protein-bound state (Foley et al., 2015; Gosai et al., 2015; Silverman et al., 2014) (**Figure 1.4B**). Unlike *in vitro* structure probing techniques, in which RNA is allowed to denature and then refold prior to probing, PIP-seq assays the native RNA structure, without denaturing and refolding steps. This technique takes tissue or cells in which RNA-protein interactions have undergone crosslinking via formaldehyde or UV light, followed by ssRNA- and dsRNA-seq in both the presence and absence of proteins, allowing for simultaneous genome-wide identification of both RNA secondary structure and RNA-protein interactions.

1.5.6 High-throughput structure probing techniques: Chemical modifiers

The desire to better understand RNA secondary structure *in vivo* lead to the development of chemical modifier-based high-throughput approaches. These chemicals can be added to tissue culture cells as well as eukaryotic organisms and modify ssRNA *in vivo*, revealing single-stranded protein unbound nucleotides. The first techniques were DMS-seq and structure-seq, both of which utilized DMS to inhibit RT progression by modifying mostly unpaired adenines and cytosines (Ding et al., 2014; Rouskin et al., 2014) (**Figure 1.5A**). Like FragSeq, these data have single nucleotide resolution, with the added advantage of being *in vivo* assays. Although DMS only modifies single-stranded nucleotides, it is worth noting that RBP binding inhibits DMS addition (Talkish et al., 2014), therefore this technique cannot differentiate between dsRNA and protein-bound ssRNA sequences.

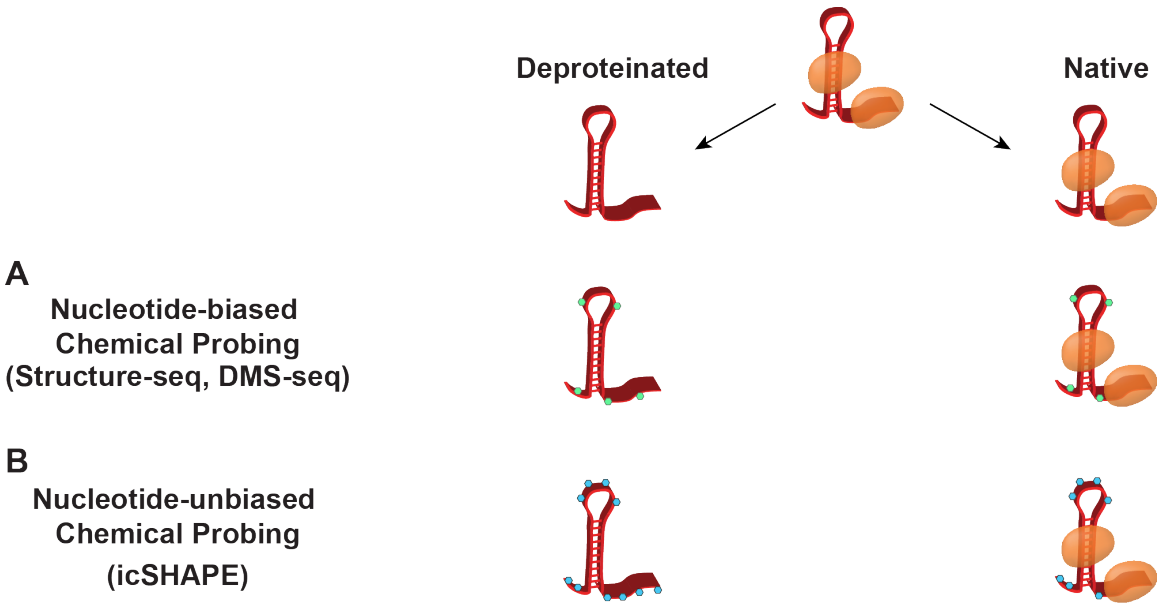


Figure 1.5: Chemical modifier-based methods for probing RNA secondary structure
A schematic representation of the chemical-based probing techniques for empirically determining secondary structure. RNA can either be probed in a native state bound by RNA binding proteins (orange ovals) or deproteinated through extraction protocols. Chemical probing works through reagents that preferentially modify nucleotides in a single-stranded confirmation, forming covalent additions in either a (A) nucleotide-biased (green hexagons) or (B) unbiased (blue hexagons) manner. While multiple covalent modifications are represented in this schematic, it is worth noting that these techniques work with single-hit stoichiometry, with one modification site interrogated per sequencing read.

A more recently developed technique is the *in vivo* click selective 2'-hydroxyl acylation and profiling experiment (icSHAPE) (Spitale et al., 2015). This method involves treatment of cells or tissues with 2-methylnicotinic acid imidazolide azide (NAI-N_3), a cell permeable chemical, which uniformly modifies the 2'-hydroxyl group of any ssRNA nucleotide (**Figure 1.5B**). The azide can then be biotinylated, allowing isolation of modified RNA with greatly reduced background, enabling SHAPE to be coupled with high-throughput sequencing. While this technique has single nucleotide resolution, it only modifies ssRNA and is subject to the same pitfalls as DMS-seq and structure-seq.

Together, next generation sequencing structure probing assays have dramatically advanced our understanding of the link between RNA secondary structure and post-transcriptional regulation. Over the last five years there has been an explosion of data, revealing fundamental features of RNA folding across various organisms.

1.6 POST-TRANSCRIPTIONAL COVALENT MODIFICATIONS

1.6.1 What are covalent modifications?

Every nucleotide of the transcriptome has the potential to undergo chemical modification. There are >100 annotated chemical modifications (Cantara et al., 2011; Limbach et al., 1994; Machnicka et al., 2013, 2013), dramatically increasing the information content of the transcriptome. These modifications have been shown to alter both RNA secondary structure and RNA-protein interactions (Arnez and Steitz, 1994; Dominissini et al., 2016; Liu et al., 2015). Most covalent modifications have been found to decrease the base pairing affinity either through interrupting the Watson-Crick edge or steric hindrance (Dominissini et al., 2016; Zhao et al., 2017; Zhou et al., 2016). Distinct secondary structures can alter the accessibility of RBP recognition sites, thereby leading to aberrant protein binding and post-transcriptional regulation (Liu et al., 2015). Additionally, a specific class of "reader" proteins is able to directly bind to these modifications, a mechanism to further regulate RBP binding (Fu et al., 2014; Xiao et al., 2016). As

covalent modifications can alter both *cis* and *trans* regulators of transcript fate, these modifications have the potential to greatly alter RNA processing in a cell type-specific manner.

Early studies of covalent modifications utilized thin layer chromatography (Davis and Allen, 1957; Desrosiers et al., 1974) and mass spectrometry (Gaston and Limbach, 2014; Meng and Limbach, 2006; Wetzel and Limbach, 2016) to identify the abundance of these modifications, unable to further study their localization or function. Since these early studies, researchers have developed biochemical techniques to identify modification sites on specific genes of interest. More recently, numerous researchers have utilized next generation sequencing techniques to generate sequencing libraries able to globally identify the sites of specific modifications transcriptome-wide. These global studies provide more in depth characterization of the modification localization allowing more in depth studies of the specific functions of different covalent RNA modifications (Dominissini et al., 2013; Fu et al., 2014), indicating roles in translation efficiency (Dominissini et al., 2016), RNA stability (Vandivier et al., 2015), and even human health (Jia et al., 2011; Zhao et al., 2014). As most studies of covalent modifications have been performed on mammalian and yeast samples, there is ample opportunity for discovery in plants.

1.7 TECHNIQUES TO STUDY COVALENT MODIFICATIONS

1.7.1 Biochemical based techniques

The first studies of covalent modifications utilized high performance liquid chromatography, electrophoresis, and thin layer chromatography in order to identify a multitude of distinct nucleotides (Davis and Allen, 1957; Desrosiers et al., 1974). Analogous to the early DNA studies identifying methyl-5-cytosine (m^5C), these initial studies found five ribonucleosides via thin layer chromatography (Davis and Allen, 1957). After the discovery of this fifth ribonucleoside, it would be over two decades before the identity of this modification was discovered (Desrosiers et al., 1974). Building on these earlier studies, researchers were able to identify noncanonical methylation events on the ribose sugar as well as on the base itself. These studies identified that

2'-O-methylation on the ribose sugar of rRNA, tRNA, and mRNA, determining that these modifications were not limited to one class of RNA (Li et al., 2005; Park et al., 2002). Unfortunately, these early techniques were limited in what information could be garnered about these modifications, unable to determine their localization along a transcript, and limited in the types of modifications that could be identified.

The covalent modification field was greatly expanded by the use of mass spectrometry to better analyze these nucleotides. In this technique, RNA molecules are digested to the single nucleotide level, then subjected to mass spectrometry. This analysis can identify changes in the molecular weight of the nucleotides, allowing an unbiased survey of the transcriptome (Gaston and Limbach, 2014; Meng and Limbach, 2006; Wetzel and Limbach, 2016). First utilized to study RNA modifications in the 1970s (Kasai et al., 1975), mass spectrometry based studies are currently being used to both identify novel modifications and quantify the prevalence of known modifications across various transcriptomes (Wetzel and Limbach, 2016). As mass spectrometry can quantitatively identify both modified and unmodified nucleotides, this analysis has linked modification to human disease. For example, N^6-methyladenosine (m^6A) has been shown to exhibit altered abundance in a variety of cancers, indicating a potential role in disease progression (Jaffrey and Kharas, 2017; Wei et al., 2017). While these early techniques are able to identify and quantify modifications, they could not determine the localization of modifications across the transcriptome.

As covalent modifications often interfere with proper nucleotide base pairing, researchers have used various enzymes to identify their localization. Covalent modifications have been shown to cause RT stalling while the enzyme attempts to insert the proper complementary base. This stalling can result in RT termination, misincorporation of a base, or proper incorporation of the complementary nucleotide (Motorin et al., 2007). RT termination at the site of specific modifications was utilized by researchers to probe modifications in specific genes of interested. To do this, primer extension was performed, and buildups of RT termination were interpreted as sites of modification (Brownlee and Cartwright, 1977; Motorin et al., 2007). The limitation to this

technique is that only a subset of modifications cause termination, while others instead result in base misincorporation or simply do not affect RT elongation. To overcome this drawback, researchers looked for chemicals that are capable of specifically modifying nucleotides of interest. One such example is N-cyclohexyl-N'-(2-morpholinoethyl)carbodiimide metho-p-toluenesulphonate (CMC). This chemical modifies all uridines in the cell, and can then be removed from unmodified uridines via alkaline hydrolysis (Ofengand and Bakin, 1997). This treatment results in an RNA population in which all pseudouridines (Ψ) are further modified by CMC, resulting in consistent RT termination. While powerful, these RT based techniques are limited to analyzing only specific genes of interest. To observe modification sites globally, several next generation sequencing techniques have been developed.

1.7.2 Transcriptome-wide identification of covalent modifications

The advent of next generation sequencing provided the first opportunity to truly characterize covalent modifications across the transcriptome. The earliest techniques developed to study modifications utilized antibodies raised against a specific modified nucleotide of interest. Reliable antibodies were raised against m^6A and were then used to perform a methyl RNA immunoprecipitation and sequencing (meRIP-seq) experiment (**Figure 1.6A**) (Dominissini et al., 2012; Meyer et al., 2012). In this technique, total or $polyA^+$ RNA is first fragmented, then incubated with the α-m^6A antibody coupled to beads. This antibody then specifically isolates m^6A-containing RNA fragments, which can be subjected to library preparation and sequencing. Comparing the meRIP library to a negative control allows peak calling algorithms to identify sequences that are enriched in the meRIP library, indicating the presence of an m^6A modification. This technique has been applied to a number of covalent modifications including N^1-methyladenosine (m^1A) (Dominissini et al., 2016; Li et al., 2016), m^5C (Hussain et al., 2013), and hydroxymethyl-5-cytosine (hm^5C) (Delatte et al., 2016). The greatest limitation to this technique is that it is unable to identify the site of modification, it is only able to result in the calling of ~100 nt peak regions in which a modification is present. This limitation has been addressed by the

development of m^{6}A individual-nucleotide-resolution crosslinking and immunoprecipitation sequencing (miCLIP-seq; **Figure 1.6B**) (Kishore et al., 2011; Linder et al., 2015). In this technique, the samples are subjected to 254 nm UV radiation during the immunoprecipitation step. This UV treatment induces mutations at the site of RNA-antibody interaction, corresponding to the m^{6}A modification. Therefore, sequencing reads with mismatches at annotated adenosine sites will correspond to the locations of m^{6}A.

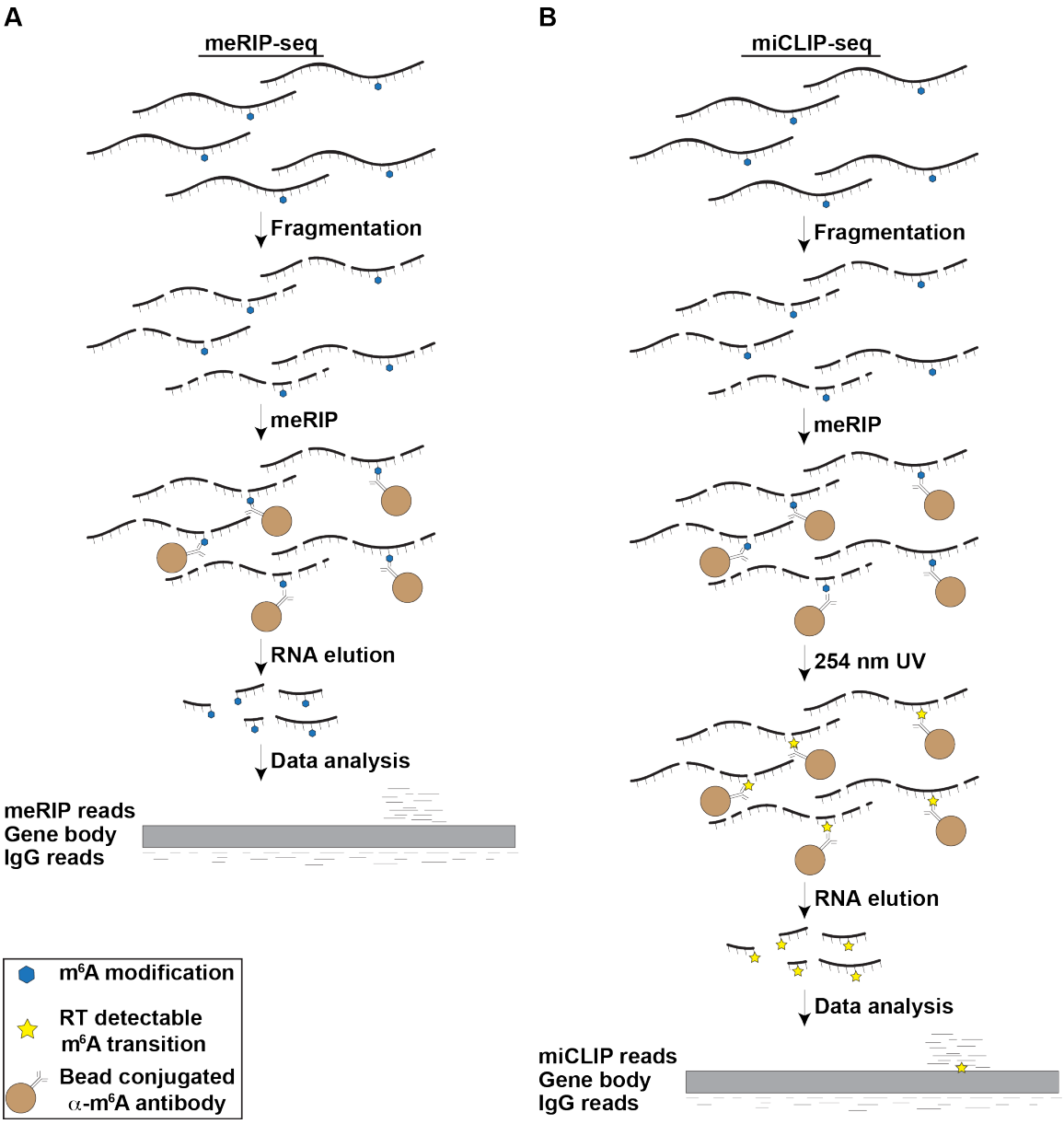


Figure 1.6: Antibody based identification of covalent modifications
(A-B) This diagram illustrates the RNA immunoprecipitation-based techniques to identify RNA modifications. In these techniques RNA is fragmented and then incubated with beads conjugated to antibodies (brown circle) to isolate fragments of modified RNA (blue hexagon). (A) In meRIP-seq the isolated RNA is then subjected to library preparation, and modification sites are identified by a buildup of sequencing reads in the meRIP sample compared to the IgG control. This identifies a large region that contains one or more modification sites. (B) In miCLIP-seq, the RNA samples bound by antibodies are subjected to 254 nm UV radiation to induce mutations at the site of modification-antibody interaction (yellow star), prior to library preparation. Modifications are then identified by a buildup of reads in the miCLIP compared to the IgG sample, with mismatches indicating the exact site of modification.

An alternative approach to identifying modifications sites utilizes the termination of RT at the sites of CMC treated Ψ to globally identify their location. Several techniques have been developed utilizing this paradigm, including Pseudo-seq (Carlile et al., 2014), Ψ-seq (Schwartz et al., 2014b), PSI-seq (Lovejoy et al., 2014), and CeU-seq (Li et al., 2015). In this technique, a sample is divided into a CMC treated and untreated control, RT is performed and terminates at the site of CMC modified Ψ, and the sample is used to generate a sequencing library (**Figure 1.7**). Buildup of read termination sites in the treated, compared to the untreated libraries, corresponds to the site of Ψ modification. While powerful, this technique is only able to identify a single modification type, whereas antibody based techniques can be used for any modification for which a highly specific antibody is available.

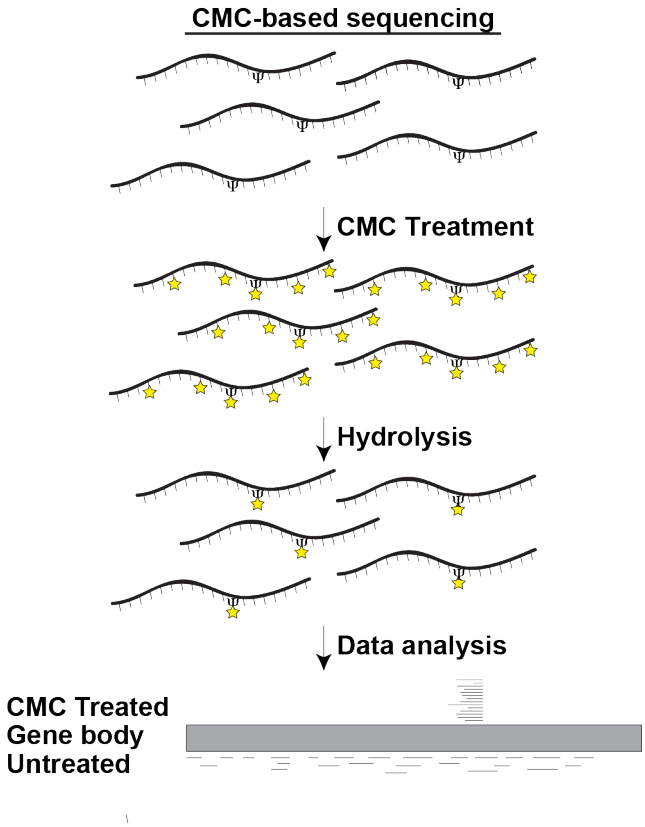


Figure 1.7: CMC-based sequencing identifies sites of pseudoruidinylation.
In CMC-based techniques, samples are first subjected to CMC treatment, which modifies all U and Ψ nucleotides in the transcriptome. A hydrolysis treatment then removes the CMC from U, leaving only Ψ modified with CMC. This treatment results in RT termination at the site of Ψ, leading to a buildup of sequencing reads terminating at the site of modification.

A third method for probing modifications utilizes the fact that modified nucleotides often result in base misincorporation by RT, rather than RT termination. The high-throughput annotation of modified ribonucleosides (HAMR) technique is an *in silico* predictor of covalent modifications that can be applied to any RNA-seq library (**Figure 1.8**). Taking known covalent modifications in yeast tRNAs, the machine learning algorithm for HAMR was used to identify the ratio of misincorporation events at various modification types (Ryvkin et al., 2013; Vandivier et al., 2015). For instance, if an annotated G has a large number of reads with Ts and As incorporated, then it often corresponds to a methyl-1-guanine (m^1G), whereas a buildup of As and Cs corresponds to a 2-methylguanosine (m^2G). HAMR has been shown to unambiguously identify Ψ, methyl-3-cytosine (m^3C), m^1G, and dihydrouridine (D). Additionally, HAMR can identify but not

distinguish between sites of isopentenyl adenine and N^6-isopentenyladenosine (i^6A | t^6A), 2-methylguanosine and 2,2-dimethylguanosine (m^2G | $m^{22}G$), or between 1-methyladenosine, 1-methylinosine, and 2-methylthio-N^6-isopentenyladenosine (m^1A | m^1I | ms^2i^6A). This allows global identification of eleven different modifications across seven distinct categories, and can be applied retroactively to any RNA-seq library. The HAMR analysis is extremely conservative, producing high false negative and true positive rates, prioritizing confident modification calls over robust identification. Therefore, this is a powerful informatics tool that will illuminate the field for years to come.

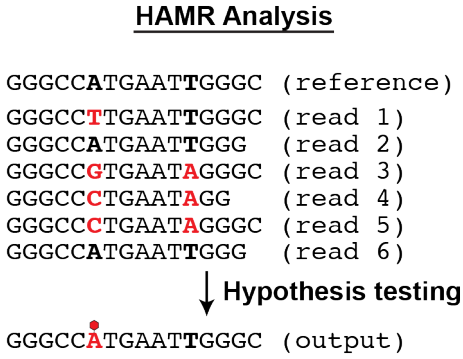


Figure 1.8: HAMR identifies modifications via *in silico* RNA-seq analysis
This figure illustrates data analysis performed by the HAMR pipeline. Six reads are mapped to the reference genome, and then searched for mismatches. HAMR excludes dinucleotide events, as these may be the product of a heterozygous sample, instead identifying tri- and tetranucleotide mismatches and testing for the presence of modifications. The output identifies the modified nucleotide, and classifies the type of modification based on the ratio of mismatches.

1.8 OUTLINE OF DISSERTATION

In Chapter 2, we utilized the PIP-seq technique in the first simultaneously study of RNA-protein interactions and RNA secondary structure across a transcriptome. We performed this analysis on the nuclei of all cells from 10-day-old *Arabidopsis* seedlings. These data revealed a global anti-correlation between RNA secondary structure and protein binding in the nuclear transcriptome. Additionally, we identified distinct protein binding and secondary structure profiles across four types of alternatively spliced exons, indicating a fundamental splicing regulatory mechanism. Lastly, we showed that using the protein-bound sequences identified by PIP-seq we could find enriched sequence motifs, indicating frequently bound sequences. Using one such

motif, we performed RNA affinity chromatography to identify a novel function of the chloroplastic RBP CHLOROPLAST PROTEIN 29A (CP29A).

In Chapter 3, we applied PIP-seq to address a specific biological question, identifying novel post-transcriptional regulators of root hair cell development. As root epidermal hair and nonhair cells are derived from the same precursor cell, we aimed to determine the role of post-transcriptional regulation in this cell fate decision. Performing PIP-seq on the nuclei or root hair and nonhair cells allowed the first comparison of global secondary structure and RNA-protein interactions between two different cell types. We identified distinct protein binding and RNA secondary structure patterns between these cell types, indicating possible mechanisms for post-transcriptional regulation. We then identified highly bound sequence motifs, and performed RNA affinity chromatography to identify highly functional RBPs. We found that the protein SERRATE (SE) functions to promote root nonhair cell fate in a microRNA biogenesis-dependent manner, while inhibiting root hair length independently of microRNA biogenesis. Additionally, we determined that the GLYCINE RICH PROTEIN 8 (GRP8) promotes hair cell fate, and functions through the phosphate starvation response pathway.

In Chapter 4, we performed RNA-seq and HAMR analysis on the nuclear and cytoplasmic transcriptomes of *Arabidopsis* primary roots and whole seedlings. We found that very few modifications present in the nucleus were also found in the cytoplasm. Additionally, we observed a distinct localization of nuclear and cytoplasmic modifications, further indicating distinct populations of covalent modifications. When probing the role of these modifications in post-transcriptional processing we found that covalent modifications present in the 5’ UTR correspond to increased mRNA stability, while those present in the CDS and 3’ UTR correspond to decreased stability relative to unmodified mRNAs. These findings indicate distinct functions for various covalent modifications.

In Chapter 5, I discuss the implications of these findings, and their impact on the field of post-transcriptional regulation, as well as future directions and remaining questions from these studies.

CHAPTER 2: GLOBAL ANALYSIS OF THE RNA-PROTEIN INTERACTION AND RNA SECONDARY STRUCTURE LANDSCAPES OF THE *ARABIDOPSIS* NUCLEUS

This section refers to work from:

- Gosai, S.J.*, **Foley, S.W.***, Wang, D., Silverman, I.M., Selamoglu, N., Nelson, A.D.L., Beilstein, M.A., Daldal, F., Deal, R.B., Gregory, B.D. (2015). Global analysis of RNA-protein interaction and RNA secondary structure landscapes of the *Arabidopsis* nucleus. *Molecular Cell. 57, 376-388.* PMID: 25557549

  *Indicates co-first author

2.1 INTRODUCTION

RNA molecules are bound throughout their lifecycle by dynamic complexes of proteins that regulate their splicing, polyadenylation, nuclear export, localization, translation, and degradation (Bailey et al., 2009). These RNA binding proteins (RBPs) interact with their targets in a sequence and secondary structure-specific manner (Cruz and Westhof, 2009). Therefore, both the bound RBPs and secondary structure are key regulatory features of these molecules (Ding et al., 2014; Li et al., 2012a, 2012b). For instance, recent studies have linked secondary structure of mRNA to translation efficiency, stability, splicing regulation, and polyadenylation (Ding et al., 2014; Li et al., 2012a, 2012b; Zheng et al., 2010).

Due to the importance of RNA secondary structure in eukaryotic post-transcriptional processing and regulation, several high-throughput approaches have been developed to globally profile single and double stranded RNAs (ssRNAs and dsRNAs, respectively) (Rouskin et al., 2014; Zheng et al., 2010). For example, ss- and dsRNA-seq employ single and double stranded RNases (ssRNases and dsRNases, respectively) to provide direct evidence for both single and double stranded regions of the transcriptome (Li et al., 2012a, 2012b; Zheng et al., 2010).

Alternatively, dimethylsulfate sequencing (DMS-seq) is a technique where samples are treated with DMS, which specifically modifies unpaired adenines (As) and cytosines (Cs) resulting in the termination of reverse transcriptase products, providing evidence for unpaired As and Cs in RNAs (Ding et al., 2014; Rouskin et al., 2014). However, recent studies have demonstrated that DMS modification is obstructed at RBP binding sites (Talkish et al., 2014), making protein-bound regions indistinguishable from truly structured regions of RNAs.

Most studies of RNA-RBP interactions identify the binding partners of a single protein of interest. This is often accomplished by crosslinking and immunoprecipitation (CLIP) (Ule et al., 2003), in which RNA-protein interactions are crosslinked via UV irradiation followed by immunoprecipitation of a protein of interest. Recently, two methods have reported development of unbiased approaches to study RNA-RBP binding (Baltz et al., 2012; Silverman et al., 2014). Protein interaction profile sequencing (PIP-seq), crosslinks RNA-protein interactions via formaldehyde, and subsequently digests ssRNA and dsRNA using structure-specific RNases before high-throughput sequencing, providing a global view of both RNA secondary structure and RBP-bound RNA sequences across the transcriptome (Silverman et al., 2014). Additionally, global photoactivatable ribonucleoside crosslinking and immunoprecipitation (gPAR-CLIP), utilizes the incorporation of a synthetic nucleotide into RNAs to identify RNA-protein crosslinking events after exposure to long wave UV radiation (Baltz et al., 2012). To date, there have been no global studies of either RBP binding or RNA secondary structure performed in the nucleus of any organism.

All aspects of post-transcriptional mRNA maturation are tightly controlled by RNA-protein interactions acting to positively or negatively regulate recruitment of catalytic molecular machines. For instance, splicing is performed by one of two large complexes, the U2- or U12-type spliceosomes which identify and excise ~170,000 or ~1,800 introns in *Arabidopsis*, respectively (Marquez et al., 2012). In addition to being regulated by multiple spliceosomes, pre-mRNA transcripts can undergo alternative splicing, resulting in mature mRNAs of different sequences (Wahl et al., 2009). In *Arabidopsis*, over 60% of introns are alternatively spliced, with failure to

excise an intron (intron retention (IR)) or exclusion of an exon (exon skipping/cassette exon (CE)) in specific isoforms comprising >64% of these events (Marquez et al., 2012). Additionally, more than 70% of *Arabidopsis* pre-mRNAs can undergo alternative polyadenylation (APA), resulting in transcript isoforms that differ in their 3' termini (Hunt et al., 2012; Wu et al., 2011). Previous studies have shown that perturbing RNA secondary structure at alternatively spliced exons can result in decreased RBP recruitment, and a shift in spliceoform abundance (Raker et al., 2009). Thus, both AS and APA are important regulatory processes driven by large collections of RBPs and their interactions with specific RNA sequences and structures.

The interplay between RBPs that bind functionally related genes has become a topic of great interest. Recent studies have attempted to identify post-transcriptional operons (Tenenbaum et al., 2011), transcripts with the same gene ontology that are bound by similar populations of RBPs. Thus, the binding of these RBPs would allow co-regulation of genes encoding functionally related proteins. Evidence for post-transcriptional operons has been observed in human cells (Silverman et al., 2014), however this analysis has yet to be performed in *Arabidopsis*.

Here, we simultaneously profile the global landscapes of RBP binding and RNA secondary structure in nuclei of 10-day-old *Arabidopsis* seedlings using our PIP-seq and structure mapping approaches. In total, this study produces the first unbiased view of RBP binding and RNA secondary structure across a nuclear transcriptome, providing a rich resource for future hypothesis generation and testing.

2.2 RESULTS AND DISCUSSION

2.2.1 PIP-seq on purified Arabidopsis seedling nuclei

To probe the RNA-RBP interaction site and RNA secondary structure landscapes of the *Arabidopsis* nucleus, we performed our PIP-seq methodology (Silverman et al., 2014) on total nuclei from 10-day-old seedlings. The nuclei were crosslinked with formaldehyde prior to purification via the isolation of nuclei in tagged cell types (INTACT) approach (Deal and Henikoff,

2010). We confirmed nuclei purity by direct imaging (**Figure 2.1A**), revealing only DAPI stained nuclei bound to the streptavidin coated beads. Additionally, we found an enrichment of the nuclear histone H3 protein and undetectable levels of the mostly cytoplasmic ACT8 (Kandasamy et al., 1999), the endoplasmic reticulum (ER)-localized BIP1 and CNX1, as well as chloroplastic RUBISCO and PEPC proteins in our INTACT-purified nuclei preparations (**Figure 2.1B**), confirming that there is no chloroplastic, ER, or cytoplasmic contamination. We used ~two million of these highly pure nuclei for each of two PIP-seq replicates, which were split into footprinting and structure only samples (four total libraries per replicate) (**Figure 2.2A**). Our structure only samples provide native structure data, and additionally serve as a background to our footprinting samples accounting for regions that are insensitive to the structure-specific RNases.

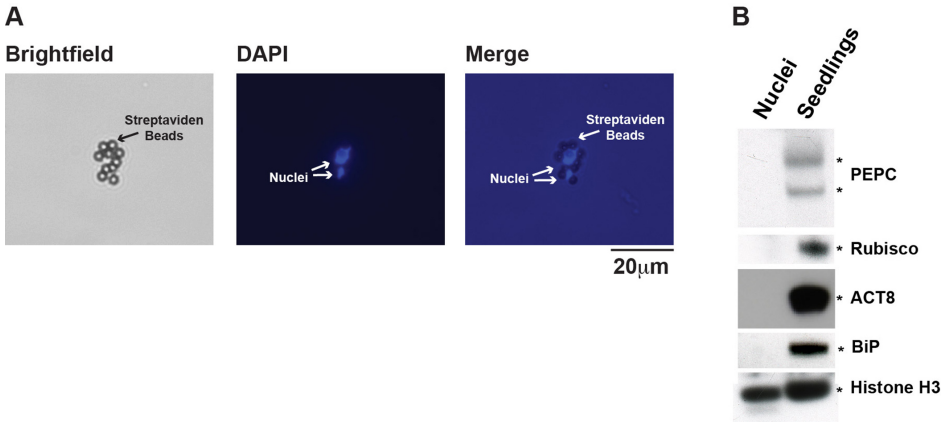


Figure 2.1: INTACT purified nuclei are free of cytoplasmic, ER, and chloroplastic contamination
(A) Microscopy imaging of DAPI stained nuclei during the INTACT purification process. The images show that only the DAPI stained nuclei are bound to the streptavidin beads. (B) Western blot of lysates from INTACT purified nuclei and 10-day-old seedlings for the chloroplastic PEPC and RUBISCO, the mostly cytoplasmic ACT8, the endoplasmic reticulum (ER)-localized BIP1 and CNX1, as well as the nuclear histone H3 proteins.

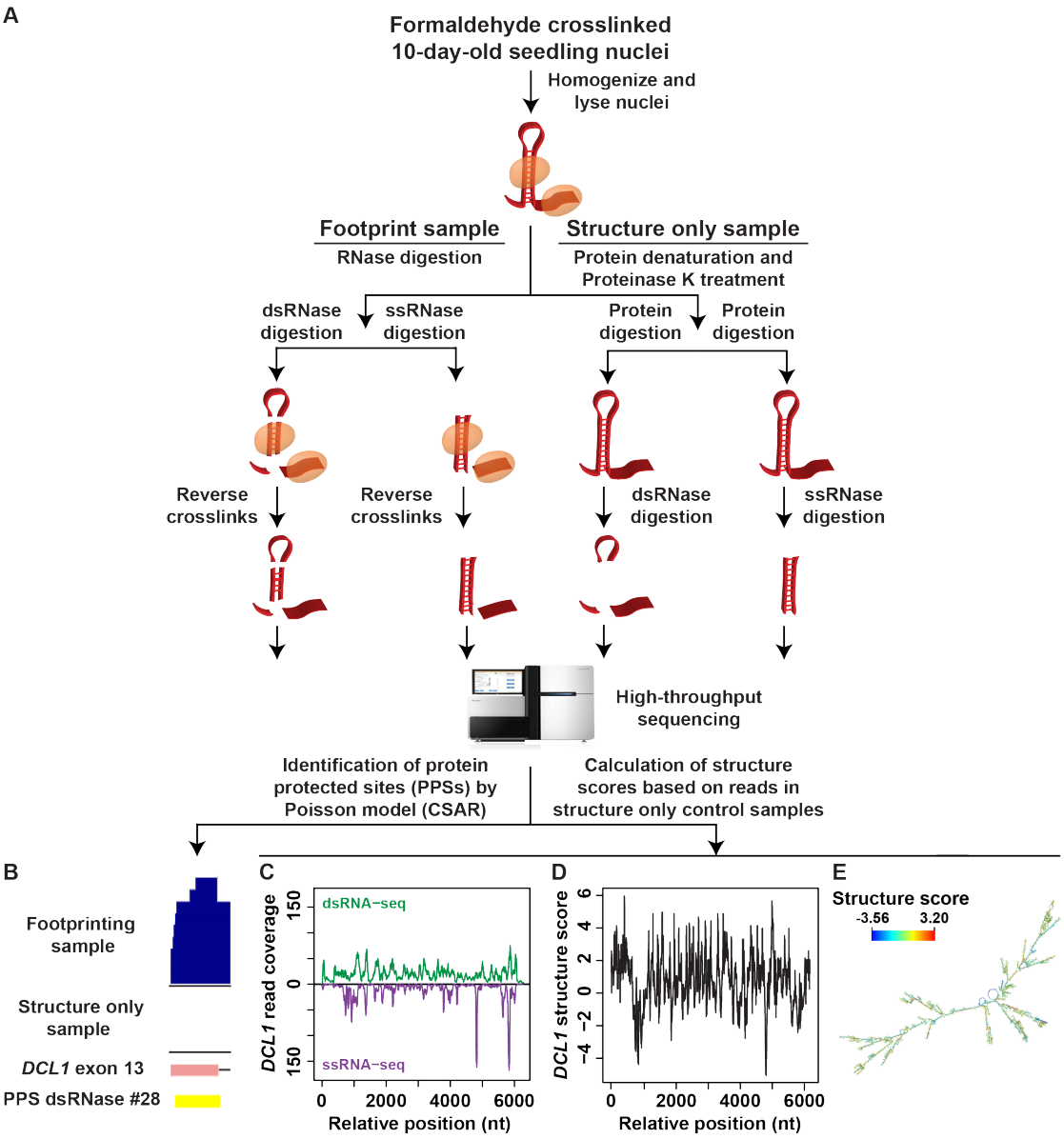


Figure 2.2: Overview of PIP-seq in *Arabidopsis* nuclei
(A) The PIP-seq approach in the *Arabidopsis* nucleus. Nuclei were purified from 10-day-old *Arabidopsis* seedlings that were crosslinked using a 1% formaldehyde solution. Nuclei were lysed and separated into footprinting and structure only samples. Four total sequencing libraries were then prepared for each replicate experiment as previously described (Silverman et al., 2014). (B) An example of PPS identification (dsRNase #28) in exon 13 of *DCL1*. (C) Read coverage across the *DCL1* transcript for the ds- (top, green line) and ssRNA-seq (bottom, purple line) structure only samples. (D) Structure scores for the *DCL1* transcript based on read coverage seen in C. (E) mRNA secondary structure model for *DCL1* determined using our methodology.

Footprint samples were directly treated with either a ss- or dsRNase. In contrast, the structure only samples were first denatured in SDS and treated with Proteinase K prior to RNase digestion. Denaturation of proteins before RNase treatment will make protein-bound sequences in the footprinting sample accessible to RNases in these reactions. Thus, RBP-bound sequences were enriched in footprinting relative to structure only samples (**Figure 2.2B**). Additionally, analysis of the structure only samples as previously described (Li et al., 2012a) allowed us to determine the native (protein-bound) RNA base-pairing probabilities for the *Arabidopsis* nuclear transcriptome (Example shown in **Figures 2.21C-E**).

The resulting high quality PIP-seq libraries (**Figures 2.3A-B**) were sequenced and provided ~24-38 million raw reads per library. To determine reproducibility, we used a 50 nucleotide (nt) sliding window to define the correlation of non-redundant sequence read abundance between biological replicates of footprinting and structure only libraries. We observed a high correlation in read counts between all footprinting and structure only libraries (Pearson correlation > 0.810) (**Figures 2.4A-D**). Similarly, principle component analysis of read coverage in 500 nt bins revealed that replicates of each library type clustered together (**Figure 2.4E**), further indicating the high quality and reproducibility of our PIP-seq libraries.

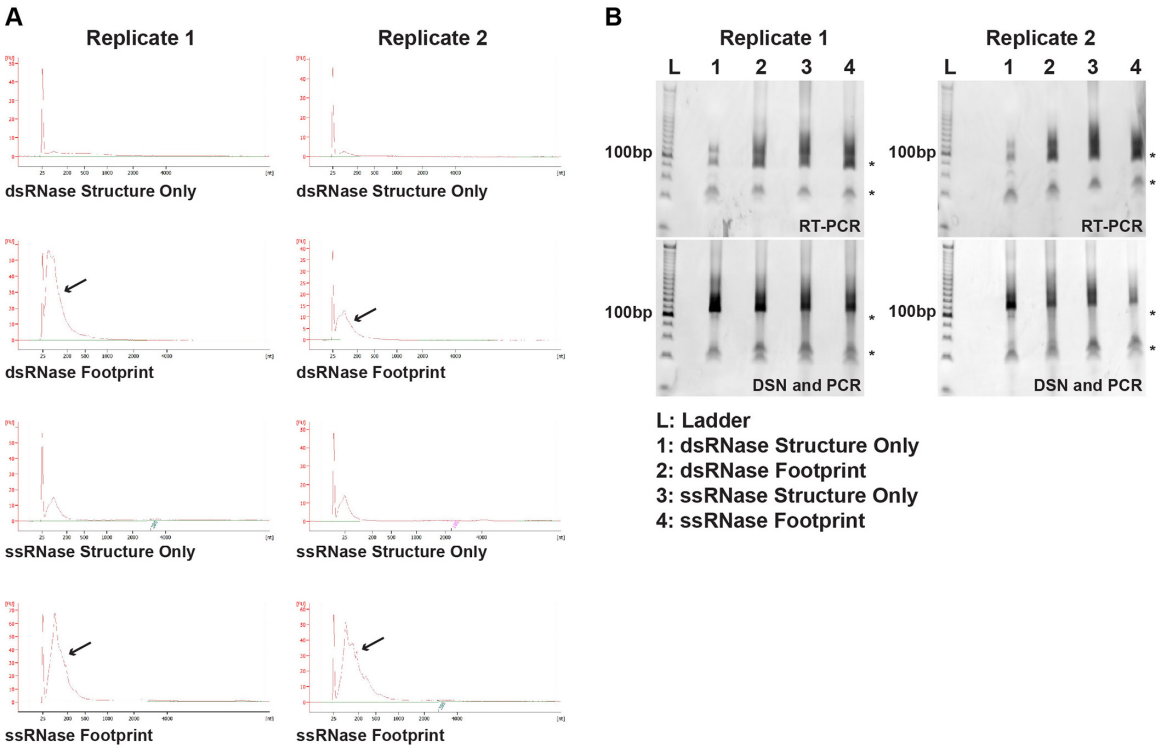


Figure 2.3: The PIP-seq libraries passed all three quality control checkpoints during library preparation

(A) Profiles from a BioAnalyzer run of the digested RNA for each of the eight PIP-seq libraries. These profiles show the expected sizes and quality for these libraries when compared to profiles from previous PIP-seq experiments. The arrows point to the larger fragments found specifically in the footprinting samples that likely represent the protein protected sites (PPSs). (B) Two size selection gels run after the initial RT-PCR or after DSN treatment and PCR. These gels show that the libraries are still of the expected high quality, and have been shifted to the expected sizes after adapter ligations. The top and bottom asterisks (*) to the right of each gel image denote adapter-adapter products and unused primers, respectively. These contaminants were avoided during the gel purification process, ensuring the high quality of our sequenced libraries.

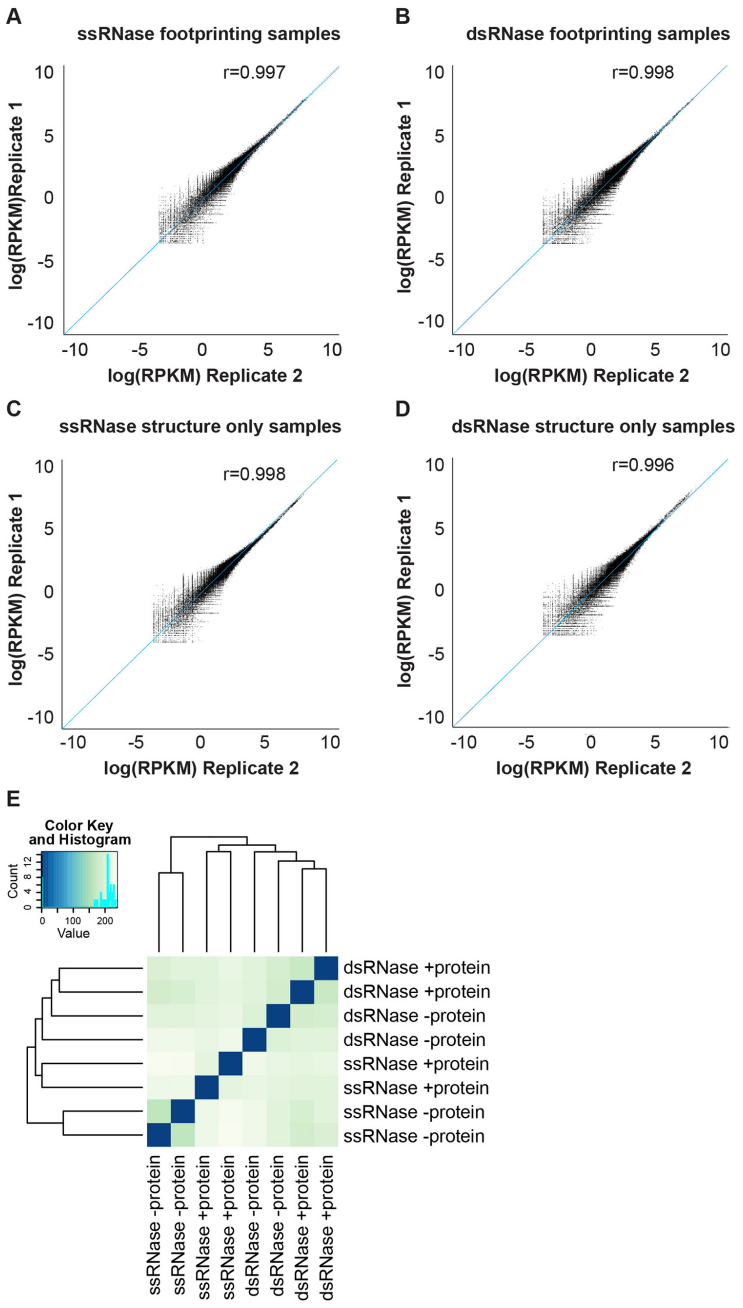


Figure 2.4: PIP-seq is a highly reproducible method
(A-B) Correlation in read counts in a 50 nt sliding window between both ssRNase (A) and dsRNase (B) footprinting replicates. (C-D) Correlation in read counts in a 50 nt sliding window between both ssRNase (C) and dsRNase (D) structure only replicates. (E) Principle component analysis of 500 nt bins between each of the eight libraries. All replicate pairs cluster together, as do both RNases demonstrating the high reproducibility of these PIP-seq libraries.

2.2.2 The RNA-protein interaction landscape of the Arabidopsis nucleus

To Identify protein protected sites (PPSs), we used a Poisson distribution model to identify enriched regions in the footprinting compared to the structure only libraries at a false discovery rate (FDR) of 5% as previously described (Silverman et al., 2014) (**Figure 2.2B**). We identified 61,632 total PPSs in our experiments, 64.7% of which overlap between the two replicates (**Figure 2.5A**), while consolidation of all PPSs yields 40,131 distinct sites. This reproducibility is much higher than many CLIP-seq experiments, which often produce <35% overlap between replicates (Lebedeva et al., 2011). The majority of PPSs were identified by the dsRNase (~30,000 PPSs) as compared to the ssRNase (~10,000 PPSs) (**Figures 2.5B-C**) treatment, with ~50% of the sites uncovered by the ssRNase overlapping those from the dsRNase libraries (**Figures 2.5D-E**).

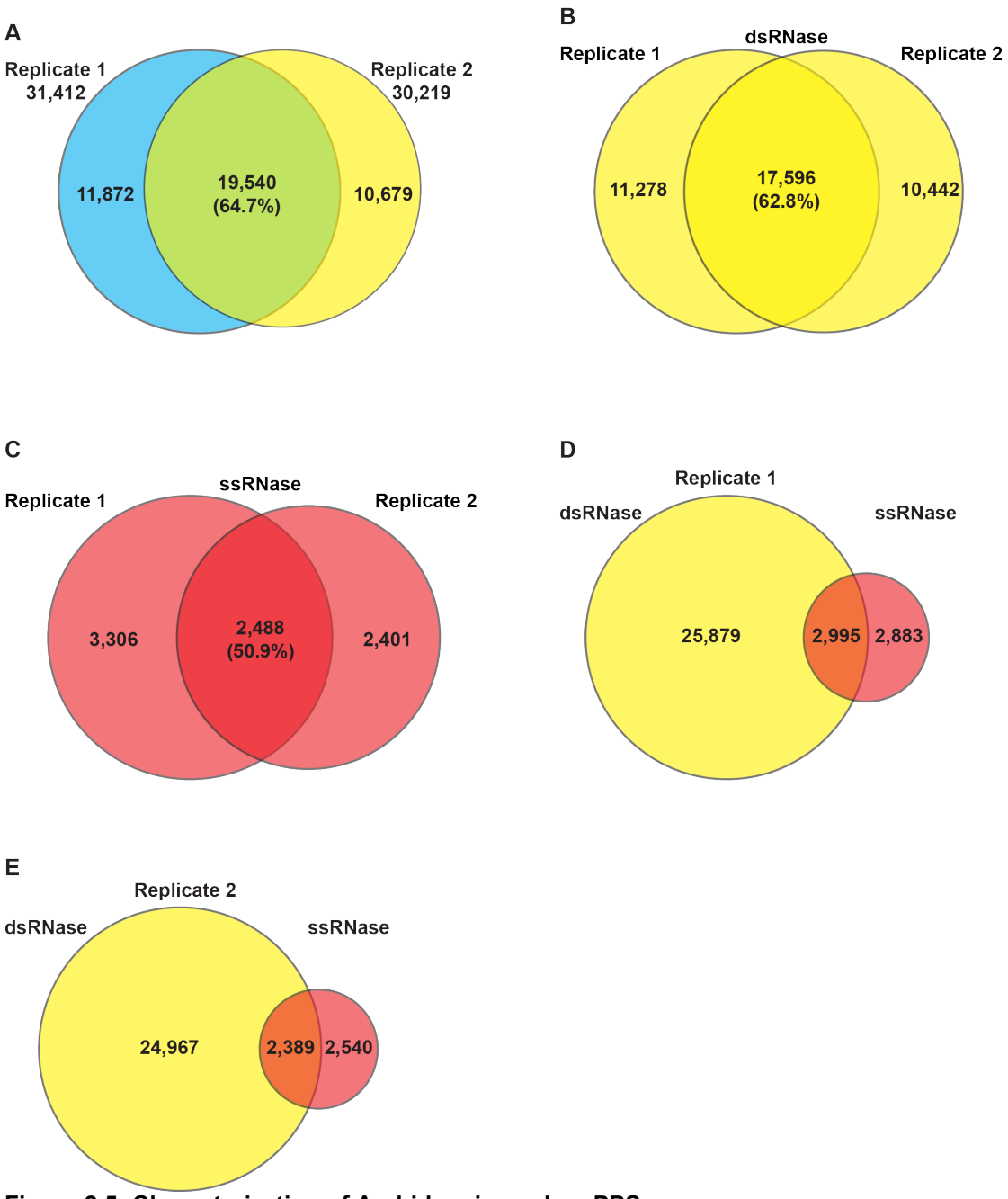
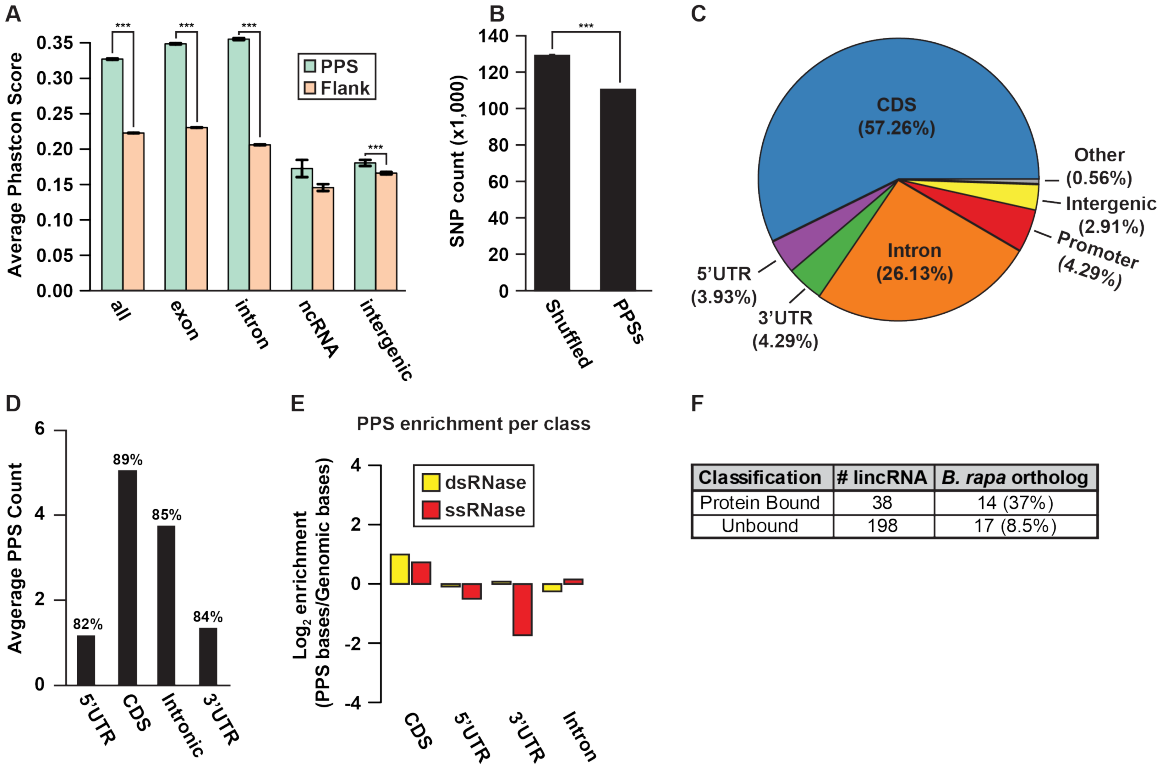


Figure 2.5: Characterization of Arabidopsis nuclear PPSs
(A) Overlap between PPSs identified from two replicate nuclear PIP-seq experiments. (B-C) Overlap in PPS calls between dsRNase- (B) and ssRNase-treated (C) PIP-seq replicates. (D-E) Overlap in PPS calls between the dsRNase- (yellow circle) and ssRNase-treated (red circle) samples for replicate 1 (D) and replicate 2 (E).

Given the high reproducibility between our PIP-seq replicates (**Figures 2.4 and 2.5**), we focused on the complete set of 40,131 distinct PPSs for all subsequent analyses. To estimate the functional relevance of these nuclear PPSs, we compared flowering plant PhastCons

conservation scores (Li et al., 2012b) for PPSs versus same-sized flanking regions. We found that PPS sequences were significantly (*p* values < $1x10^{-200}$; χ^2 test) more evolutionarily conserved than flanking regions (**Figures 2.6A**). Importantly, this was true for PPS sequences in both exonic and intronic portions of the nuclear collection of mature and pre-mRNA transcripts (nuclear mRNAs), but not for ncRNAs (**Figure 2.6A**). These results support the notion that nuclear mRNA sequences are constrained by their ability to interact with RBPs, while decreased PPS conservation within ncRNAs is consistent with their low conservation rates across plant species (Liu et al., 2012).



Classification	# lincRNA	*B. rapa* ortholog
Protein Bound	38	14 (37%)
Unbound	198	17 (8.5%)

Figure 2.6: Characterization of Arabidopsis nuclear PPSs
(A) Comparison of average PhastCons scores between PPSs (green bars) and equal sized flanking regions (orange bars) for various genomic regions. *** denotes p-value < $1x10^{-10}$, Fisher's t-test. (B) Analysis of the total number of SNPs identified by the 1001 Genomes Project (Cao et al., 2011) in PPSs compared to a shuffled background control. *** denotes p-value < $1x10^{-10}$, χ^2 test. (C) Absolute distribution of PPSs throughout various RNA species and transcript regions. (D) Average PPS count per pre-mRNA transcript region. Percentages indicate the fraction of annotated RNAs that contain sequencing information for that region. (E) Genomic enrichment of PPS density, measured as $\log_2$ enrichment of the fraction of PPS base coverage normalized to the fraction of genomic bases covered by indicated nuclear mRNA regions for the dsRNase- (yellow bars) and ssRNase-treated (red bars) libraries. (F) Breakdown of bound

compared to unbound nuclear lincRNAs that are conserved between *Arabidopsis thaliana* and *Brassica rapa*.

We also reasoned that functional RBP interacting sequences would contain less nucleotide diversity across closely related strains when compared to an equal number of same-sized regions randomly selected from detected transcripts. To address this, we used data from the 1001 Genome Project, which has catalogued naturally occurring single nucleotide polymorphisms (SNPs) between eighty strains of *Arabidopsis thaliana* (Cao et al., 2011). We found a significant (p value < 2.2×10^{-16}; χ^2 test) decrease in nucleotide diversity within PPSs compared to shuffled regions (**Figure 2.6B**). Therefore, *Arabidopsis* PPSs resist the effects of random genetic drift occurring in the numerous populations across the globe, indicating their functional significance.

A classification of all distinct PPSs revealed the majority of these sites were located in nuclear mRNAs, with the largest fractions occupying the coding sequence (CDS) (57.3%) and introns (26.1%) (**Figure 2.6C**). Closer examination of PPSs broken down by genic features (e.g. 5' and 3' UTR, CDS, and intron) revealed that detected *Arabidopsis* nuclear mRNAs contained multiple binding events in both the CDS (~5 total/gene) and introns (~4 total/gene), while the 5' and 3' UTRs averaged only a single interaction per expressed transcript (**Figure 2.6D**).

We then tested the enrichment of PPSs in specific nuclear mRNA regions (e.g. 3' and 5' UTRs) normalized to the number of bases annotated as these features in the TAIR10 *Arabidopsis* genome. We found that PPSs identified by both RNases were enriched in CDSs, while being underrepresented in 5' UTRs (**Figure 2.6E**). Interestingly, both introns and 3' UTRs show opposite enrichment trends for ds- and ssRNase-treated samples, suggesting that PPSs preferentially occur in more highly or lowly structured regions respectively. In total, our results reveal that the CDSs of mRNAs are enriched for RBP binding in the *Arabidopsis* nucleus.

Although PPSs in ncRNAs were not conserved, this category consists of many RNA subgroups, thus conserved classes might be obscured. Long intergenic noncoding RNAs (lincRNAs) are a recently discovered class of ncRNAs that are necessary for vertebrate development (Cech and Steitz, 2014; Sauvageau et al., 2013), but are not well characterized in

plants (Hacisuleyman et al., 2014; Liu et al., 2012). We examined the relationship between our PIP-seq data and a set of ~2,700 curated lincRNAs in *Arabidopsis* (Liu et al., 2012) to identify nuclear protein-bound RNAs. We detected 236 lincRNAs in our nuclear sequencing data, 38 of which contained one to four PPSs (**Figure 2.6F**). We found that these protein-bound lincRNAs were significantly (p value < 4.5×10^{-30}; χ^2 test) more conserved within the related crop species *Brassica rapa* (37%, 14 total) as compared to unbound nuclear lincRNAs (8.5%, 17 total) (**Figure 2.6E**). The combination of nuclear protein binding and conservation in *B. rapa* suggests that RBP-bound nuclear lincRNAs have important functions in plant systems.

2.2.3 Patterns of RNA secondary structure and RBP binding are anti-correlated

To address the overall landscape of RBP binding and RNA secondary structure in specific regions of nuclear mRNAs, we calculated the structure scores and PPS densities and examined the average profiles for all detectable transcripts. The structure score is a generalized log-ratio of dsRNA-seq to ssRNA-seq reads at each nucleotide position, with positive and negative scores indicating ds- and ssRNA, respectively. To examine the relationship between PPS density and structure score, we focused on the boundaries between the UTRs and CDS of nuclear mRNAs. We observed the highest PPS density in the CDS with decreased occupancy within the 5' and 3' UTRs (**Figures 2.7A-B**), consistent with the gross PPS localization and enrichment analysis (**Figures 2.6C-E**). Interestingly, we observed significantly (p value < 8.2×10^{-32}; Wilcoxon test) higher levels of protein binding directly over the start codon (**Figure 2.7A**) relative to immediate flanking regions. Similarly, we examined the start codons at high confidence upstream open reading frames (uORFs) (von Arnim et al., 2014) and found a significant (p value < 0.01; Wilcoxon test) increase in PPS density over uORF start codons relative to the upstream flanking region (**Figure 2.7C**). Similar increases in PPS density over the start and stop codon were speculated to be due to ribosome binding. However, the nuclear preparations used in this study are free of the cellular compartments containing functional ribosomes (cytoplasm and ER) (**Figure 2.1B**) and RBP binding profiles for transcripts that are not translated in the rough ER

(**Figure 2.7D**) demonstrate the same protein binding profile. Taken together, these results suggest that one or more nuclear RBPs occupy this region before export to the cytoplasm.

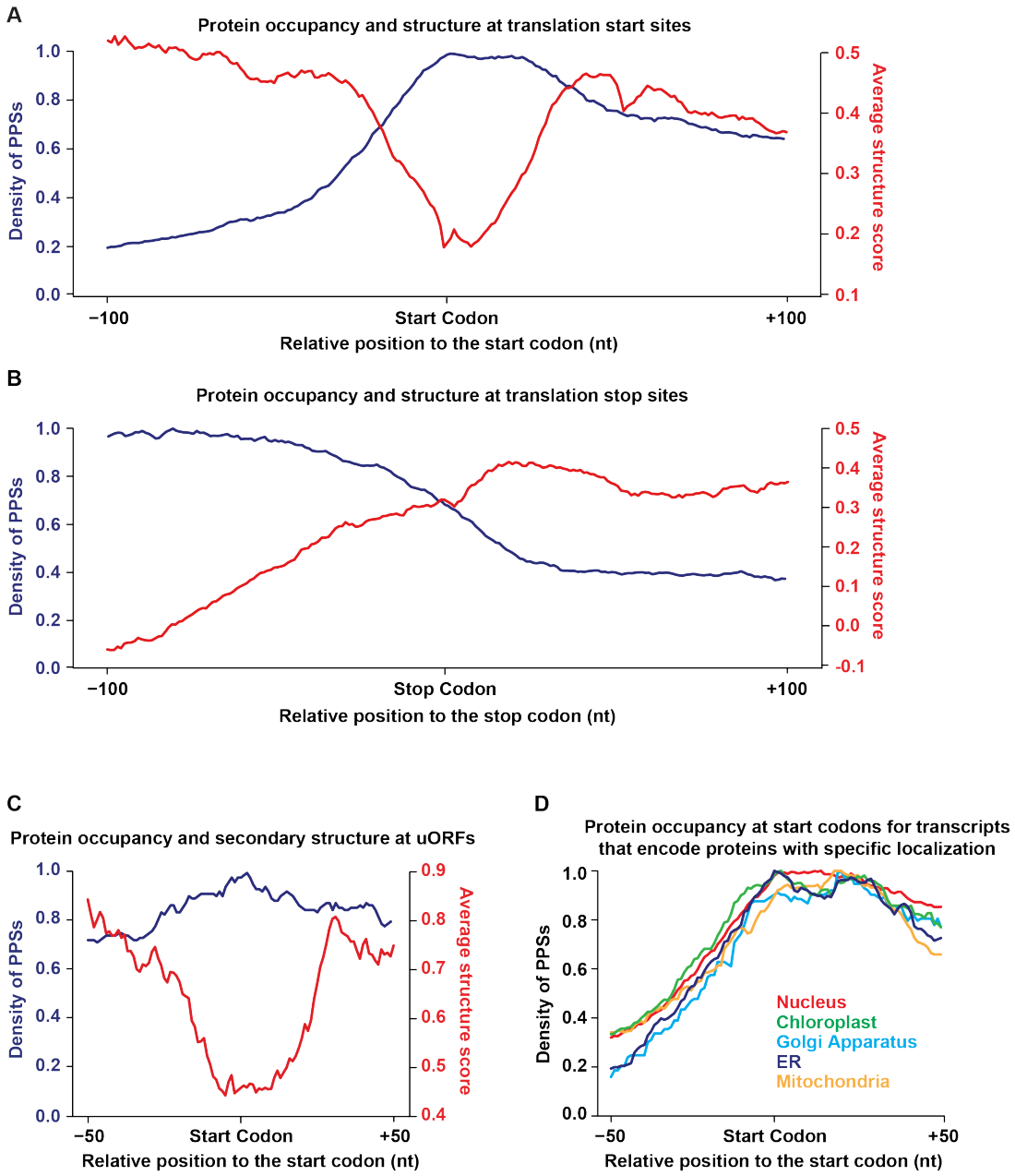


Figure 2.7: Patterns of protein occupancy and secondary structure at mRNA start and stop codons

(A-B) PPS density and structure score profiles for nuclear mRNAs based on our PIP-seq experiments. Average PPS density (blue lines) and structure scores (red lines) at each position +/- 100 nt from canonical (A) start and (B) stop codons for *Arabidopsis* nuclear mRNAs. (C) PPS density and structure score profiles for highly confident uORFs (von Arnim et al., 2014). Average PPS density (blue) and structure score (red) at each position +/- 50 nt at uORF start codons. (D)

Average PPS density at each position +/- 50 nt at canonical start codons for transcripts encoding proteins that are localized to specific cellular compartments (as specified by colored line and label) based on TAIR10 annotation.

In contrast to RBP occupancy, we found that secondary structure was higher in both UTRs compared to the CDS at the regions analyzed, with a significant (*p* values < 0.05; Wilcoxon test) dip directly over uORF and canonical start codons, as well as upstream of the stop codon, as observed previously (Ding et al., 2014; Li et al., 2012b) (**Figures 2.7A-C**). Thus, these structural characteristics at the start and stop codons seem to be a consistent feature of both *Arabidopsis* nuclear and mature mRNAs. Interestingly, our analyses revealed that secondary structure and PPS density are anti-correlated to one another. Specifically, we looked at both PPS density and structure score simultaneously, and found a significant (*p* value < 2.2×10^{-16}; asymptotic *t* approximation) anti-correlation (Spearman's rho < -0.82) between these metrics at both canonical start and stop codons. Although the correlation is milder (likely due to fewer instances), there is a significant (*p* value < 3.6×10^{-9}; asymptotic *t* approximation) negative correlation (Spearman's rho < -0.55) for uORF start codons as well.

It is worth noting that although the majority of PPSs were identified in the dsRNase-treated samples this does not necessitate that the interacting RBPs are binding dsRNA. In support of this hypothesis, we found that more highly structured regions generally surrounded PPSs, with a lower average structure score directly over the RBP-bound region (**Figure 2.8A**). Although the dsRNase identified PPSs have a significantly (*p* value < 2.2×10^{-16}; Wilcoxon test) higher average structure score than those uncovered by the ssRNase (**Figure 2.8A**), the dip in structure score directly over these regions suggests that they can be ds- or ssRNAs. Taken together, these results suggest that many *Arabidopsis* RBPs bind ssRNA flanked by highly structured regions.

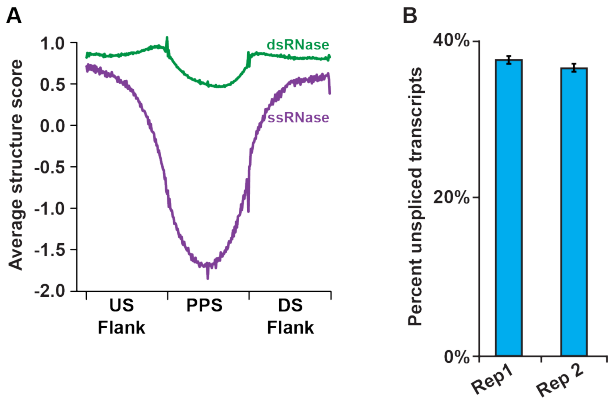


Figure 2.8: Secondary structure and protein binding landscapes at protein interaction sites and isolated alternative splicing events.
(A) The average structure score of exonic PPSs from the dsRNase (green) or ssRNase (purple) treated libraries, and equal sized flanking regions, for 100 equal sized bins. (B) The mean percentage of exon/intron junction mapping reads per transcript from two replicates (as indicated) of total RNA sequencing for congruently purified nuclei. Error bars represent standard error of the mean (SEM).

It should also be noted that the higher overall structure of the UTRs compared to the CDS is opposite to what has been observed previously both *in vivo* and *in vitro* when profiling total (mostly mature cytoplasmic) RNA in *Arabidopsis* (Ding et al., 2014; Li et al., 2012b). Together, these results suggest that the structural landscape of the nucleus is distinct from that of the cytoplasm. These differences in secondary structure in specific cellular locales will need to be further investigated.

As we were probing the nuclear transcriptome, we next examined the PPS density and structure scores across all TAIR10 annotated splice donor and acceptor sites. We first determined that the RNA population consisted of a high percentage of unspliced pre-mRNA. Specifically, we found that ~42% of reads mapping to the first and last constitutively spliced intron junctions cross the exon-intron boundary in total RNA sequencing datasets from congruently purified nuclei, suggesting comparable levels of spliced and unspliced transcripts in our datasets (**Figure 2.8B**). Despite the large percentage of detectable unspliced transcripts (pre-mRNAs), structural and PPS profiles across exonic and intronic regions cannot be directly compared due to lower read coverage in introns. Therefore, we first compared 30 nt regions up- or downstream of acceptor and donor intron sites, respectively, and found that the 3' end of introns had significantly

(p value < $1x10^{-30}$; Wilcoxon test) higher protein binding relative to the 5' end (**Figure 2.9A**). These results are consistent with the U2 auxiliary factors (U2AFs) occupying the acceptor splice site (Wahl et al., 2009). Intriguingly, there were distinct patterns of secondary structure at both the splice donor and acceptor sites (**Figure 2.9A**). Upstream of the donor site, we observed a dramatic decrease in secondary structure from nucleotides -3 to -1, corresponding to the U1 snRNA binding site (-3 to +8) (Chiou et al., 2013). This dip in secondary structure mirrors what we have seen over the translation start codon (**Figure 2.7A**), revealing that this region is more accessible to intermolecular RNA pairing than flanking sequences, perhaps facilitating binding of the U1 snRNA. Additionally, we found a drop in secondary structure immediately upstream of the splice acceptor site, suggesting an increased accessibility to U2AFs and other splicing factors in this region (Wahl et al., 2009) (**Figure 2.9B**).

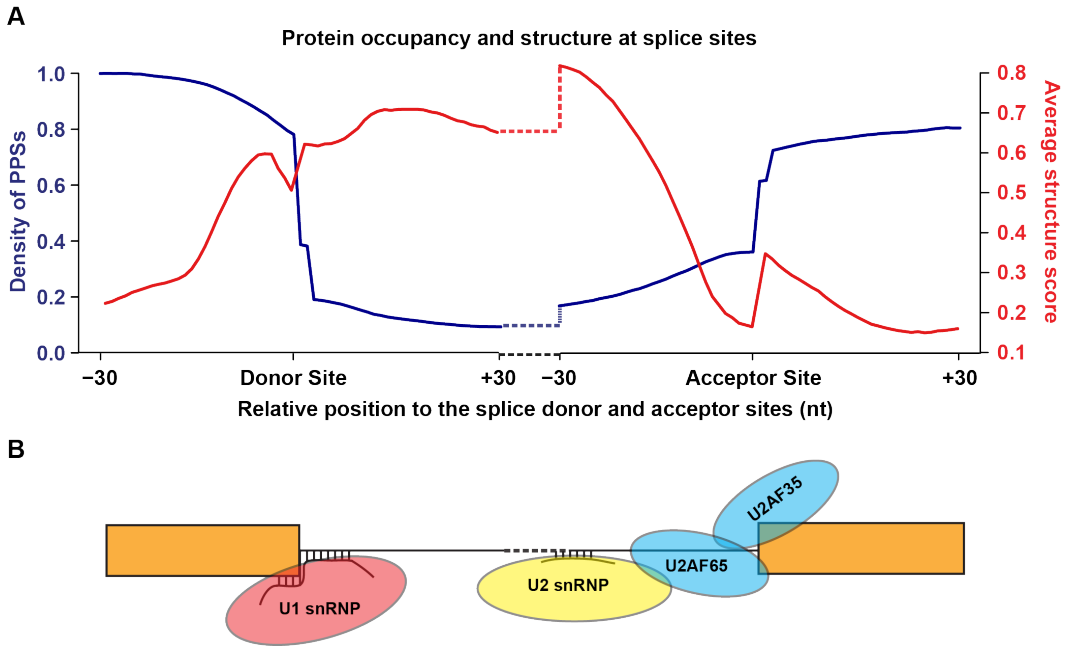


Figure 2.9: Patterns of protein occupancy and secondary structure at pre-mRNA splice sites
(A) PPS density and structure score profiles for exon/intron boundaries of nuclear mRNAs. Average PPS density (blue lines) and structure scores (red lines) at each position +/- 30 nt from splice donor and acceptor sites. (B) Model depicting the canonical protein and RNA interactions of the U2-type spliceosome at the splice donor and acceptor sites depicted in A.

We again observed opposing patterns of secondary structure and PPS density at all regions examined in these analyses (**Figure 2.9A**). Specifically, we found that this anti-correlation

(Spearman's rho < -0.93) between PPS density and RNA secondary structure was significant (*p* value < $2.2x10^{-16}$; asymptotic *t* approximation) at regions flanking the acceptor sites, as well as the upstream exonic sequence at donor sites. The proximal intronic region at donor sites had a milder (Spearman's rho < -0.38) but still significant (*p* value < 0.05; asymptotic *t* approximation) anti-correlation between structure score and PPS density, which may be due to the intermolecular base pairing between the U1 snRNA and the intron (**Figure 2.9B**) that occurs at eight of the 30 nt probed. In total, our findings reveal that RBP binding and RNA secondary structure are anti-correlated features in the *Arabidopsis* nuclear transcriptome.

2.2.4 Distinct RNA secondary structure and RBP binding profiles demarcate alternative splicing and polyadenylation

The specific patterns of RBP binding and RNA secondary structure at exon/intron boundaries suggest that these features may also have distinct distributions at sites of AS. Therefore, we compared the profiles for these two features at several types of alternatively spliced exons. To do this, we used ASTALAVISTA (Foissac and Sammeth, 2007) to annotate alternative splicing events in the TAIR10 transcript assembly, and isolated all examples of CE and IR events. We also focused on TAIR10 introns that have been previously described as U12-dependent splice sites (Marquez et al., 2012). We compared average PPS density and structure score for 50 nt in the exonic region and 30 nt in the intronic sequence at both the splice donor and acceptor sites for these splicing events (**Figure 2.10A**). We found that IR events have significantly (*p* values < $4.3x10^{-7}$; Wilcoxon test) higher PPS density in the 40 nt upstream (-40 to -1) of the splicing donor, while CE and U12-type introns do not significantly (p-value > 0.05, Wilcoxon test) differ from constitutive introns. This trend for increased PPS density continues in IR events 30 nt into the intron at splice donor sites, with these events showing ~4.5 fold higher protein binding than constitutive introns (*p* value < $1.9x10^{-44}$; Wilcoxon test) (**Figure 2.10B**). The increased binding within these introns is consistent with the presence of intronic splicing silencers (ISS), *cis* elements that recruit proteins to inhibit spliceosome assembly (Chen and Manley,

2009). We observed increased PPS density at the splicing acceptor for both CE and IR sites in the downstream exon (p-values < 6.7×10^{-6}, Wilcoxon test) and in the 30 nt of intron directly upstream of this splice site (p-values < 0.001, Wilcoxon test) (**Figure 2.10B**). This can likely be explained by recruitment of RBPs through a combination of both positive and negative *cis* regulatory elements, such as exonic splicing silencers (ESS) to induce exon skipping, and intronic splicing enhancers (ISE) to increase inclusion, working additively to regulate each exon in a cell type-specific manner (Chen and Manley, 2009). In total, these results reveal that IR and CE events can be differentiated from one another based on the patterns of protein binding density just up- and downstream of both splice sites.

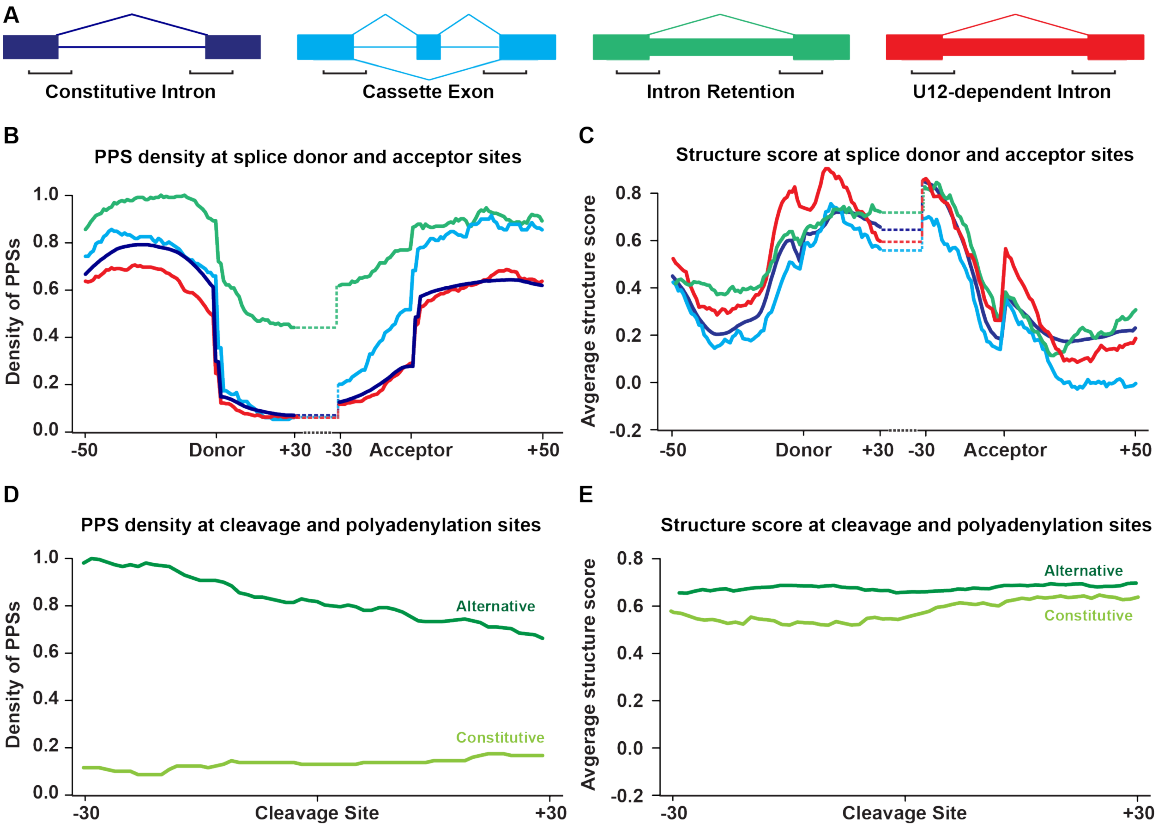


Figure 2.10: Protein occupancy and secondary structure landscapes at alternative splicing and polyadenylation sites.
(A) Diagram of constitutive introns (blue), cassette exons (turquoise), intron retention events (green), and U12-type introns (red). Large boxes represent exons, lines represent constitutive introns, and small boxes represent alternatively spliced introns, with the black brackets indicating

the regions graphed in B and C for reference. (B) PPS density profiles for constitutive and alternative splicing events in *Arabidopsis*. Average PPS density at each position -50 to +30 nt at the donor splice site, and -30 to +50 at the acceptor splice site. Line colors correspond to examples shown in A. (C) Structure score profiles for constitutive and alternative splicing events in *Arabidopsis* covering the same regions as B. Line colors correspond to examples shown in A. (D) PPS density profiles for constitutive and alternative poly(A) sites of pre-mRNAs. Average PPS density at each position +/- 30 nt from constitutive (light green line) and alternative (dark green line) cleavage and polyadenylation sites. (E) Average structure score profiles for constitutive and APA sites covering the same regions as D.

We next probed the structural profiles for each of these subsets of introns across splice sites (**Figure 2.10C**). The most striking feature we observed was the dramatic difference in overall profile shape between U12-type introns and constitutive introns upstream of the donor splice site (-16 to -1). We found a significantly (*p* value < 0.01; Wilcoxon test) higher structure score for the introns in this region, which have a PPS profile that is indistinguishable from constitutive introns. This structural profile likely influences the identity of the proteins binding this region (Cech and Steitz, 2014), resulting in distinct RBP populations at each type of intron. Additionally, IR events are also significantly (*p* value < 4.5×10^{-3}; Wilcoxon test) more structured 40 nt upstream of the donor splice site (-40 to -1). Specifically, these profiles reveal highly structured regions that are associated with increased binding levels of regulatory proteins. Thus, in both U12-dependent and IR events the increased structure in specific regions likely limits the accessibility of binding sites to specific RBPs allowing for a tighter control over the splicing machinery. Interestingly, CEs are the only subset of events that is consistently less structured than constitutive introns. This trend is only statistically significant (*p* value < 0.05; Wilcoxon test) upstream of the acceptor site (-30 to -1), but the analysis is limited by a low number of annotated events (< 700) (**Figure 2.10C**). In total, these results reveal that each of these three subtypes of alternative splicing events has a distinct combination of PPS and structural profiles, supporting the idea that both structure and protein occupancy are required for their proper regulation.

Polyadenylation, the addition of the poly(A) tail during eukaryotic mRNA maturation, is a highly regulated event. Therefore, we calculated average PPS density and structure score 30 nt up- and downstream of expressed transcripts with constitutive or APA sites (Sherstnev et al., 2012). We found that APA events were on average 3.7-fold (*p* value < 4.8×10^{-16}; Wilcoxon test)

more protein-bound up- and downstream of the cleavage site as compared to constitutive events (**Figure 2.10D**). Interestingly, there is no significant (*p* value > 0.05; Wilcoxon test) difference in structure scores between the alternative and constitutive sites (**Figure 2.10E**), revealing that this differential protein binding is independent of secondary structure. These results indicate that APA sites do not exhibit altered secondary structure compared to constitutive sites, however the increased protein binding could be used to differentiate these two types of events from one another.

2.2.5 The structural landscape of protein-bound RNA motifs

To identify RBP-bound motifs, we employed the motif finding algorithms MEME (Bailey et al., 2009) and HOMER (Heinz et al., 2010) on PPSs partitioned by specific region (e.g. CDS, intron, or UTRs) or on the entire collection, respectively. We identified one GAN repeat motif by MEME that was common to both the CDS and 5' UTR (**Figure 2.11A**), while HOMER identified 40 octamers that were significantly (E-values < 10^{-7}) enriched in our PPSs, of which we further characterized four of the most significantly enriched (E-values < 1.0×10^{-67}) (**Figures 2.11B-E**).

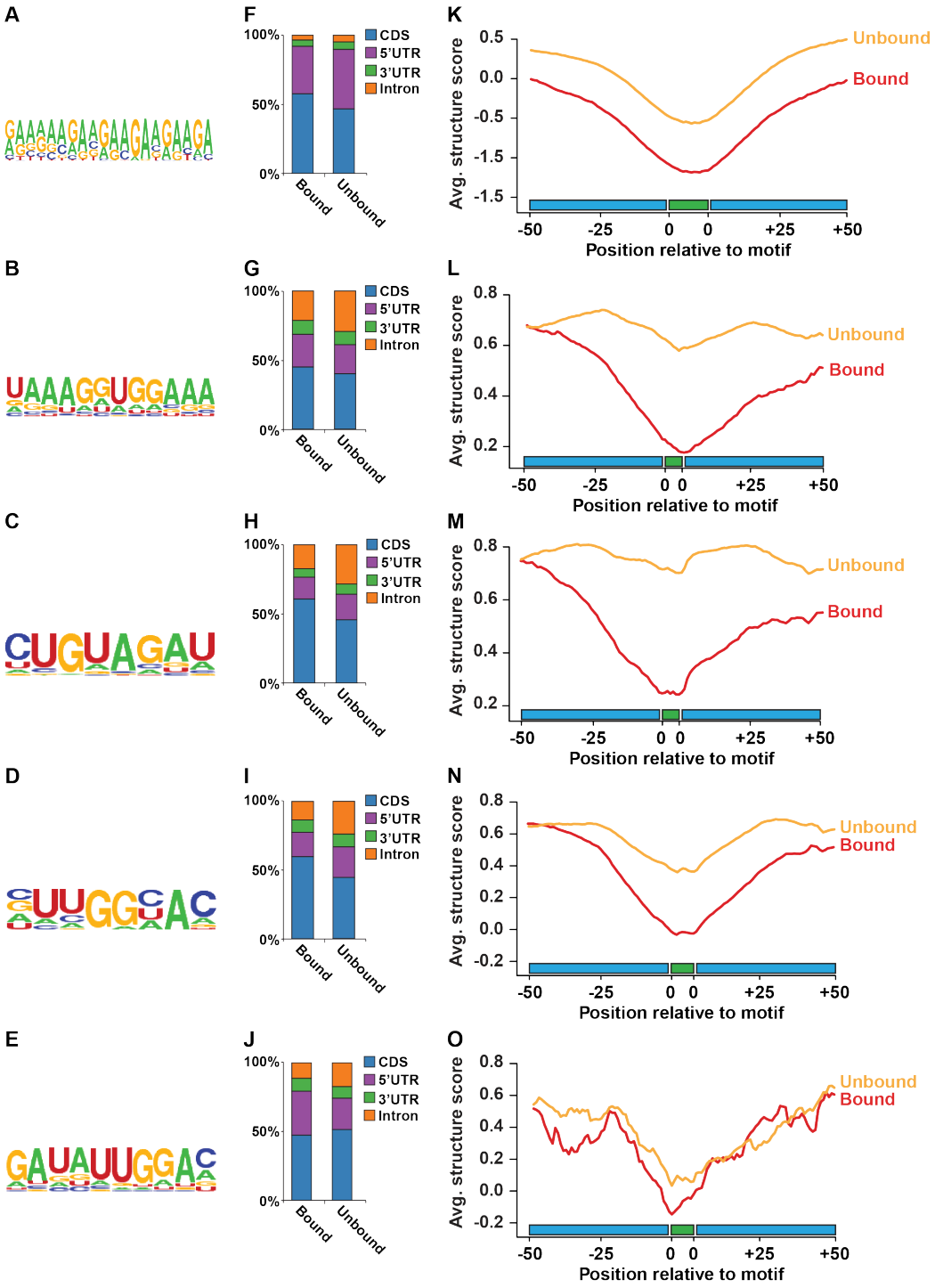


Figure 2.11: The landscape of protein-bound RNA motifs
(A-E) Overrepresented sequence motifs identified by MEME (A) or HOMER (B-E) analysis of PPS sequences. (F-J) The relative distribution of protein-bound and unbound motifs from A-E throughout specific pre-mRNA regions, including the CDS (blue), 5' UTR (purple), 3' UTR (green), and intron (orange). (K-O) Structure score profiles at RBP binding motifs from A-E. Average

structure score at each position +/- 50 nt up- and downstream of bound (red lines) and unbound (orange lines) motif occurrences from A-E.

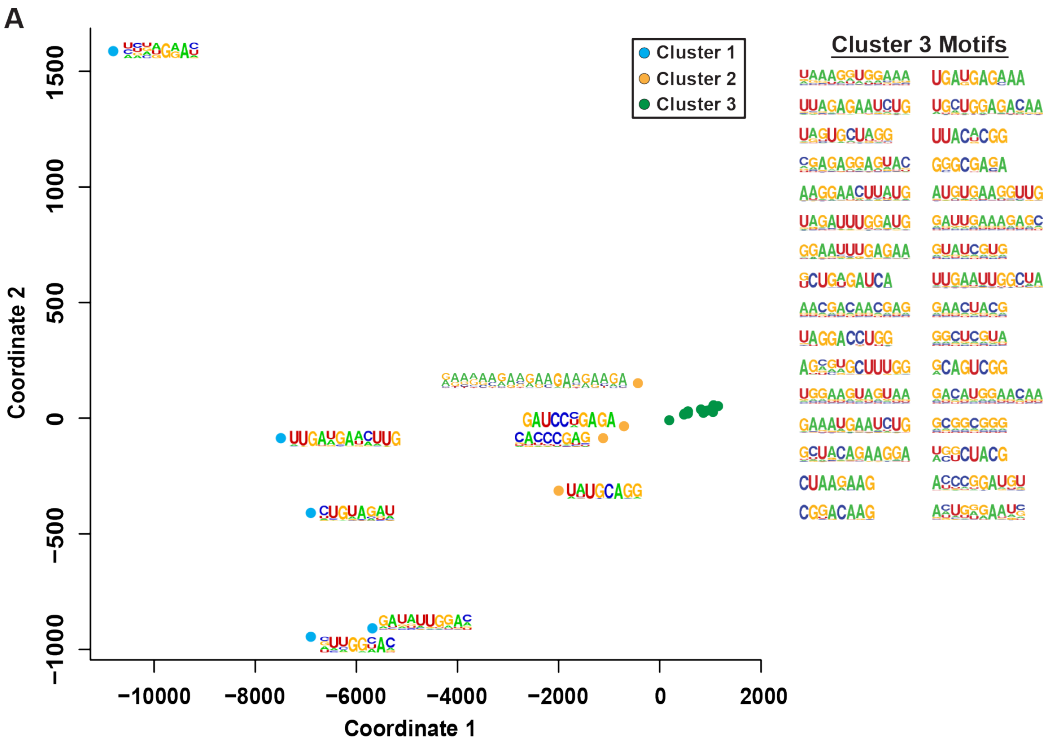
We identified the percentage of PPS-bound and -unbound motif occurrences in specific regions of nuclear mRNAs normalized by their overall length in the genome (**Figures 2.11F-J**). Comparing the localization of bound and unbound motif instances revealed stark differences. We saw an overall enrichment of bound sites within the CDS and 5' UTR. Conversely, the unbound HOMER motif instances were generally more prevalent in introns (**Figures 2.11G-J**), while the 5' UTR is overrepresented in the unbound GAN repeat occurrences (**Figure 2.11F**). In total, these results indicate that within the nucleus RBP binding is enriched within 5' UTR and CDS instances of specific sequence motifs.

To define the structural context at these five enriched motifs, we calculated average structure scores at the core motif and 50 nt flanking regions for bound and unbound instances. We observed that the five motifs have low structure scores, but are flanked by more structured regions (**Figures 2.11K-O**). As mentioned above, the high levels of this conformation within the nuclear transcriptome may explain increased PPS identification by the dsRNase (**Figures 2.5 and 2.8A**). Interestingly, protein-bound instances of all five motifs and their flanking sequences are significantly (p values < 7.3×10^{-12}; Wilcoxon test) less structured relative to unbound instances of these sequences (**Figures 2.11K-O**). In total, these findings support the observations that PPSs occur preferentially at less structured regions of transcripts. Whether this is a cause or consequence of protein binding to these sequence elements will need to be further investigated.

2.2.6 Evidence of post-transcriptional operons in the Arabidopsis nuclear transcriptome

RBP interacting motifs often co-occur in functionally related genes in human cells (Silverman et al., 2014), but it is not known if this happens in the *Arabidopsis* nuclear transcriptome. To address this, we interrogated the interactions between protein-bound motifs discovered by our PIP-seq approach. Thus, we identified all bound instances of each identified motif in target RNAs using the HOMER suite (Bailey et al., 2009) on the total set of nuclear PPSs.

We then quantified co-occurrences of each pair of these protein-bound motifs within all nuclear mRNAs. We used k-means clustering of the resultant weighted adjacency matrix and identified three clusters of motifs that co-occur on highly similar sets of target transcripts (**Figure 2.12A**). Interestingly, Clusters 1 and 2 have only five and four motifs respectively, while Cluster 3 consisted of the remaining 32 motifs, although no transcripts contained more than four of these co-occurring PPS-bound motifs. The number of transcripts containing at least three bound motifs within each cluster varied greatly, with Clusters 2 and 3 having 188 and 204 transcripts respectively, while Cluster 1 had the most co-occurring bound motifs with 5,887. These findings indicate that many *Arabidopsis* transcripts contain numerous RBP interacting motifs.



B

Cluster #	GO Annotation	P-value
Cluster 1	Post-embryonic development	2.5×10^{-64}
	Organ development	4.2×10^{-17}
	Cell death/apoptosis	1.2×10^{-13}
Cluster 2	Response to desiccation	2.8×10^{-10}
	Response to abscisic acid	1.1×10^{-5}
	Response to cold	2.2×10^{-5}
Cluster 3	Stress response	2.8×10^{-8}
	Post-translational modification	1.1×10^{-5}
	mRNA processing	1.8×10^{-5}

Figure 2.12: Clusters of motifs are present in functionally related genes
(A) Multidimensional scaling (MDS) analysis of RBP-bound motif co-occurrence in *Arabidopsis* transcripts. The motifs used for this analysis were identified by HOMER- and MEME-based analyses of PPS sequences. Colored dots indicate cluster membership as defined by k-means clustering (k = 3). (B) The most significantly enriched biological processes (and corresponding *p* value) for target transcripts of the specified clusters of motifs identified in A where three or more of the motifs are protein-bound and co-occurring.

We used agriGO (Du et al., 2010) to interrogate over-represented biological processes for these collections of RNAs with co-occurring RBP-bound motifs (**Figure 2.12A**). We found that the most highly over-represented functional terms were related to distinct processes, including

cell death/apoptosis, post-embryonic, and organ development (Cluster 1); response to desiccation, abscisic acid, and cold (Cluster 2); as well as stress response, post-translational modification, and mRNA processing (Cluster 3) (**Figure 2.12B**). The identification of groups of functionally related transcripts bound by the same collection of RBPs during their nuclear life cycle supports the idea of post-transcriptional operons (Keene and Tenenbaum, 2002; Tenenbaum et al., 2011) functioning in the *Arabidopsis* nucleus.

2.2.7 CP29A localizes to the Arabidopsis nucleus

After identifying enriched motifs within our PPS list we used these motifs to identify putative *Arabidopsis* RBPs. To begin, we confirmed that these sequences interact *in vitro* with specific RBPs using a UV crosslinking assay with radiolabeled RNA probes (from **Figures 2.11A-E**), or a scrambled control sequence. We found that each sequence motif interacted with one or more distinct RBPs (**Figure 2.13A**). We then used these same probes in RNA-affinity chromatography assay followed by mass spectrometric. Using this approach with four significant HOMER motifs (**Figures 2.11B-E**), we identified 25 proteins with peptides that were enriched over our negative controls (**Figure 2.13B**), with four proteins that passed a threshold of >6 fold enrichment for interaction with at least one sequence (**Figure 2.14A**). Interestingly, CVP2 as well as the LRR family and DUF544 containing proteins do not have canonical RNA binding domains (RBDs). This is similar to recent findings in human RBP identification (Baltz et al., 2012; Castello et al., 2012), suggesting that these proteins interact with their target motifs via non-canonical RBDs or an RBP partner.

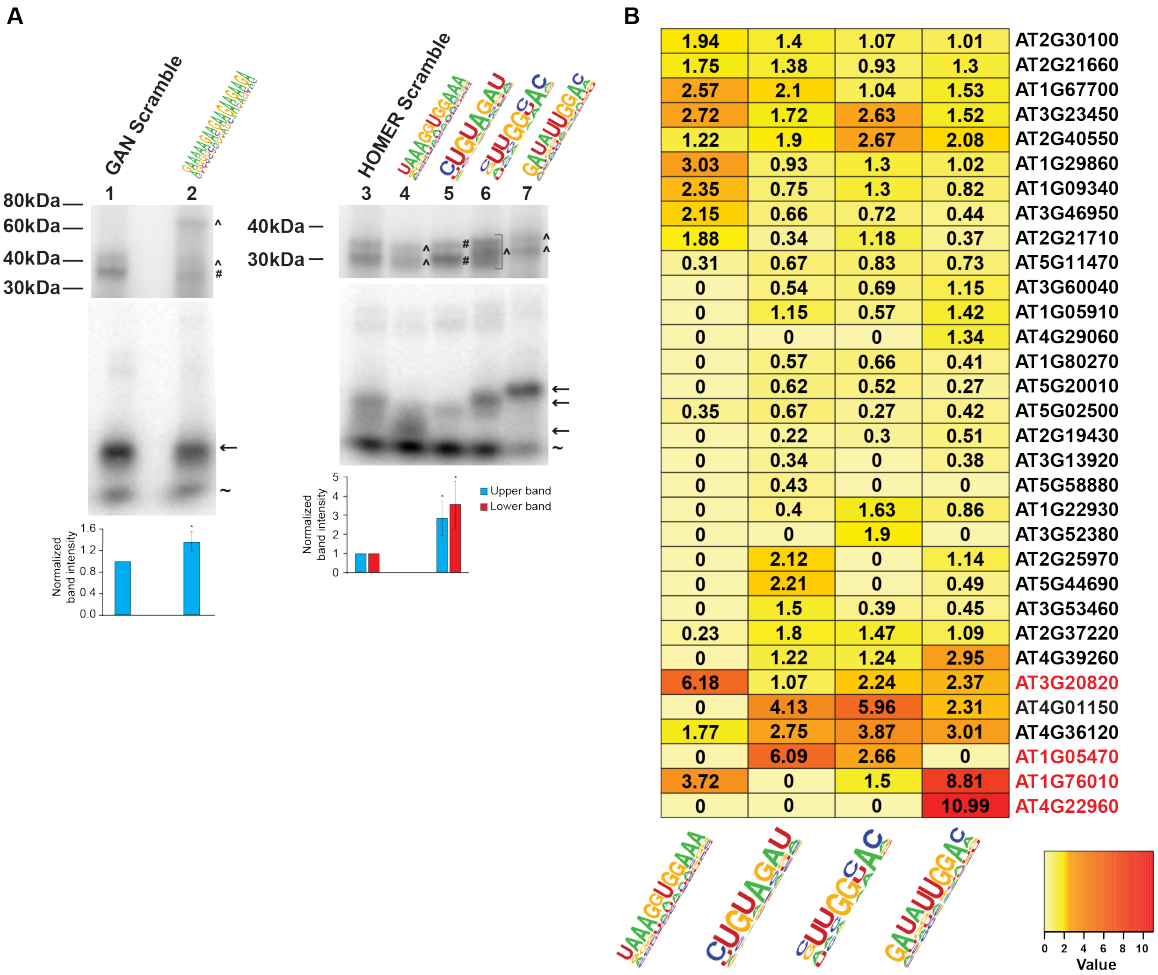


Figure 2.13: Identification of putative RBPs using synthetic RNA motifs
(A) UV-crosslinking analysis for the indicated RBP-interacting motifs compared to non-specific controls using *Arabidopsis* 4-week-old leaf lysate. Three biological replicates were performed, and a representative gel is shown. For bands that are present in both the motif and scrambled control lanes, the intensity of the band was quantified and normalized to the unbound probe, and is graphed in the below chart, as fold change relative to scrambled control. ^ denotes bands that are present in a motif lane, but are absent from the scrambled controls. # denotes a band that is present in both the motif lane and the scrambles control, and therefore was quantified in the graphs below the respective lane. The arrows denote unbound probes. ~ denotes the unincorporated radiolabeled ATP. * denotes *p* value < 0.05, Fisher's t-test. Error bars represent standard deviation, n=3. (B) Enrichment of peptides from the indicated *Arabidopsis* proteins as compared to negative control pulldown samples. The number of peptide spectrum matches (PSM) for each sample was taken and the percentage of the total PSM for each identified protein was calculated. The fold change relative to the average of the empty bead and scramble bait negative controls was then graphed. Proteins denoted in red are candidate RBPs that passed a 6-fold enrichment threshold.

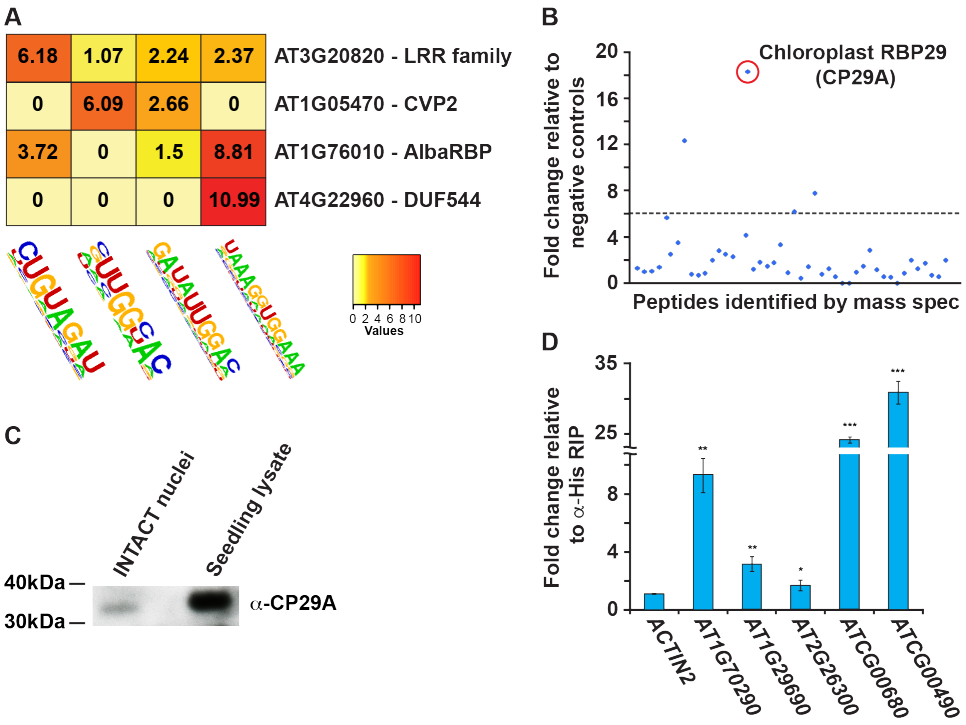


Figure 2.14: Identification of Arabidopsis RNA interacting proteins
(A-B) Identification of proteins that interact with specific over-represented sequence motifs. (A) The fold enrichment over negative control of peptides from each designated protein identified by mass spectrometry analysis of eluates after RNA-affinity chromatography with each specified motif. (B) The fold enrichment of peptides from proteins identified by mass spec analysis after RNA-affinity chromatography with the GAN repeat motif (**Figure 2.11A**). The top candidate identified by this analysis, CP29A, is annotated and denoted with a red circle. Dotted line indicates cutoff of 6-fold enrichment. (C) Western blot analysis of INTACT-purified nuclei and *Arabidopsis* 10-day-old seedling lysates using an antibody to CP29A. (D) RT-qPCR analysis of three nuclear GAN motif containing genes (*AT1G70290*, *AT1G29690*, and *AT2G26300*), two positive control chloroplast transcripts (*ATCG00680* and *ATCG00490* [also with motif]), and an *ACTIN2* negative control following RIP with an α-CP29A or α-His antibody. The data is presented as the fold change in the α-CP29A relative to α-His RIP samples. Error bars, ±SD. *, **, and *** indicate *p* value < 0.05, 0.001, and $1x10^{-10}$, respectively.

The GAN repeat motif is of particular interest because it has been linked to splicing regulation in *Physcomitrella patens* (Wu et al., 2014). The UV crosslinking assay indicated that numerous proteins were capable of binding this motif, with several 25-40 kDa proteins significantly (*p* value < 0.05; Fisher's t-test) enriched over the negative control (**Figure 2.13A**). However, from mass spectrometry analysis of interacting proteins only four passed a threshold of 6-fold enrichment over negative controls, with the strongest candidate being CHLOROPLAST RNA BINDING PROTEIN 29 (CP29A; AT3G53460; >16-fold enrichment) (**Figure 2.14B**). This

protein has previously been identified as a RBP that functions in the chloroplast (Ye et al., 1991), but nuclear localization had not been demonstrated. We used an *Arabidopsis* CP29A monoclonal antibody (Kupsch et al., 2012) to perform western blots on lysates from INTACT-purified nuclei and 10-day-old seedlings. Although at low levels, we could reproducibly detect CP29A in the *Arabidopsis* nucleus (**Figure 2.14C**), in contrast to other chloroplastic proteins (**Figure 2.1B**), showing that a subset of CP29A is localized in the nucleus.

To confirm that CP29A could interact with both nuclear and chloroplast transcripts containing the predicted GAN repeat motif *in vivo,* we performed RNA immunoprecipitation (RIP). We used lysates from formaldehyde-treated leaves and incubated them with either a monoclonal α-CP29A or α-His antibody (negative control) followed by RT-qPCR for three nuclear transcripts and two chloroplast RNAs as positive controls. All three nuclear and one chloroplast (*ATCG00490*) transcript contain the GAN repeat motif. We found that all five transcripts were significantly (all *p* values < 0.05; Fisher's t-test) enriched >1.5 fold in the α-CP29A compared to the α-His control RIP samples, as opposed to the *ACTIN2 (*AT3G18780*)* negative control (**Figure 2.14D**). Taken together, these results indicate that CP29A localizes to both the chloroplast and nucleus, and interacts with a subset of GAN repeat motif containing transcripts in *Arabidopsis*, suggesting a new functionality for this plant RBP.

2.3 CONCLUSION

Here, we characterized the global landscapes of RNA secondary structure and RBP occupancy of the *Arabidopsis* nuclear transcriptome (**Figure 2.2**). We demonstrated that these data are highly reproducible and that the identified protein binding sites are significantly more conserved than their flanking sequences (**Figure 2.6A**). Additionally, we calculated the structure score for nuclear RNAs that passed filtering criteria, creating a comprehensive database of *in vivo* RNA secondary structure for the *Arabidopsis* nucleus (**Figures 2.2C-E**). Together, these datasets provide a vast resource of RBP binding and secondary structure information for the *Arabidopsis*

nuclear transcriptome that can inform future experiments focused on understanding post-transcriptional regulation.

Using the data generated here, we searched for patterns of global RBP binding and RNA secondary structure. The most striking association that we identified was a distinct anti-correlation between RNA secondary structure and RBP occupancy within the RNA regions that were examined (**Figures 2.7A-C, 2.9A, 2.10, and 2.11**). This pattern was present when focusing on uORF and canonical translation start codons (**Figures 2.7A and 2.7C**), stop codons (**Figure 2.7B**), exon/intron junctions (**Figure 2.9A**), and specific RBP binding motifs (**Figures 2.11K-O**). Furthermore, we found that the RBP-interacting motifs identified by our study tend to be less structured when protein-bound (**Figures 2.11K-O**). Although we cannot discern causality, our findings reveal that in general RBPs bind to unstructured sequence elements in target transcripts resulting in the overall opposing patterns of these features in the *Arabidopsis* nucleus.

When initially examining these data we questioned whether the structure score was artificially lowered in regions of high PPS density by occlusion of the RNase through the incomplete digestion of bound RBPs. However, if this were true these regions would not be called as PPSs in our initial analyses because their read levels would be artificially raised in the structure only libraries as well. Furthermore, we find that the presence of PPSs is actually associated with more negative structure scores (**Figures 2.8A and 2.11K**). Thus, our results are likely true biological observations of decreased structure at RBP binding sites, not an artifact of the PIP-seq methodology.

We also examined subsets of annotated alternative exons and identified unique profiles of PPS density and secondary structure in constitutive, CE, IR, and U12-type introns (**Figures 2.10B-C**). These profiles suggest that gross protein binding can regulate alternative splicing, while secondary structure can influence the population of proteins that occupies each region. Although it is known that RBP binding in the exon or intron can regulate alternative splicing (Chen and Manley, 2009; Simpson et al., 2010), this is the first observations of protein occupancy levels in regions near the splice site differentiating subsets of alternative exons. Our observations have

provided the resources for identifying these populations of proteins and specific structural features in these alternative events.

Finally, we uncovered motifs that were enriched within our PPSs and identified co-occurrences of RBP-bound instances of these sequences in functionally related transcripts (**Figure 2.12**). These findings are similar to previous observations in human cells (Silverman et al., 2014), and support a model in which RNA transcripts encoding proteins with related functions also share a set of interacting RBPs through underlying sequence motifs allowing their co-regulation. Taken together, our findings suggest that both plants and humans use different groups of RBPs to allow specific sets of proteins, especially those functioning in development, stress responses, and apoptosis, to be precisely co-regulated in an operon-like fashion.

CHAPTER 3: A GLOBAL VIEW OF RNA-PROTEIN INTERACTIONS REVEALS NOVEL ROOT HAIR CELL FATE REGULATORS

This section refers to work from:

- Foley, S.W., Gosai, S.J., Wang, D., Selamoglu, N., Solitti, A.C., Köster, T., Steffen, A., Lyons, E., Daldal, F., Garcia, B.A., Staiger, D., Deal, R.B., Gregory, B.D. (2017) A global view of RNA-protein interactions reveals novel root hair cell fate regulators. *Dev. Cell.* In Press

3.1 INTRODUCTION

The agricultural industry is responsible for providing food for an ever-growing global population. Currently, population growth is on track to outpace agricultural growth by the year 2050 (OECD and FAO, 2012). This challenge is compounded by climate change, which reduces arable land that can be used for agricultural production, necessitating the development of new technologies to increase crop yield. One method to achieve this is through the study of plant root development, as roots function in the uptake of both water and nutrients from the environment (Grierson et al., 2014). Thus, these studies can result in the engineering of plants that can better tolerate and respond to these environmental stresses, without affecting the development of the agriculturally important aerial tissues.

The plant root epidermis is responsible for absorbing both water and nutrients from the environment (Grierson et al., 2014). During root growth, epidermal precursor cells differentiate (Dolan et al., 1993) into either root hair or nonhair cells. The long hair-like projections of hair cells dramatically increase surface area, allowing uptake of more nutrients from the surrounding soil. Therefore, plants regulate the ratio of root hair to nonhair cells in a manner that is partially dependent on environmental signals (Meisner and Karnok, 1991). More specifically, plants grown under nutrient or water poor conditions develop more hair cells with longer hairs (Bates and

Lynch, 1996), thereby greatly increasing the surface area of the root to promote increased absorption.

Phosphate limitation is one of the most common nutrient stresses that plants face when growing in fields for agriculture production. This is because roots can only absorb inorganic phosphates, which are naturally present at very low concentrations in soil (Patrick and Khalid, 1974). Therefore, plants have developed numerous mechanisms by which to maximize the uptake of this nutrient in phosphate poor soil (Niu et al., 2013; Williamson et al., 2001). In fact, researchers have described three major changes in *Arabidopsis thaliana* (hereafter *Arabidopsis*) root development during phosphate starvation. First, the primary root ceases downward growth, with a subsequent increase in lateral roots branching away from primary roots (Linkohr et al., 2002). Additionally, the root epidermis dramatically increases the number of root hair cells, while also increasing hair length (Bates and Lynch, 1996). Finally, root epidermal cells secrete acid phosphatases, enzymes that catalyze the conversion of organic into inorganic phosphates, which can be subsequently absorbed (Gilbert et al., 1999). Thus, there is a clear link between response to phosphate starvation and root hair cell fate. However, the molecular mechanisms by which exogenous phosphate levels regulate this cell fate decision are not fully understood.

To better understand this cell fate decision, previous studies have focused primarily on understanding the transcriptional networks present in both hair and nonhair cells (Lee and Schiefelbein, 2002). Two key transcription factors that function in this process are WEREWOLF (WER) and CAPRICE (CPC), which promote nonhair cell (Ryu et al., 2005), and hair cell fate, respectively (Wada et al., 1997). Plants that have null mutations in these genes have dramatic root epidermal phenotypes. However, hair and nonhair cells are never fully absent (Lee and Schiefelbein, 2002). The presence of both cell types, even when these key transcription factors are absent, suggests that there are other pathways that regulate root hair cell fate. In fact, more recent studies have begun to appreciate the numerous post-transcriptional processes that may influence this cell fate decision. Specifically, a recent study identified hair cell-specific alternative

splicing events (Lan et al., 2013), indicating splicing as one potential post-transcriptional mechanism of cell fate decision regulation.

Beginning with its transcription, each RNA molecule is bound by an ever-changing cohort of RNA binding proteins (RBPs). These proteins regulate RNA stability, post-transcriptional processing (capping, splicing, etc.), export, localization, and translation (Jangi and Sharp, 2014; Vandivier et al., 2016; Younis et al., 2013). Furthermore, a single RBP can bind to and potentially regulate the transcripts encoded by thousands of different genes (Cruz and Westhof, 2009), allowing these proteins to act as master regulators of developmental switches (Han et al., 2013; Warzecha et al., 2009). However, whether RBPs regulate *Arabidopsis* root hair cell fate decisions and development is currently unknown.

Like transcription factors, RBPs bind to primary sequence motifs. However, the intricate secondary structures that each RNA molecule forms adds an additional mechanism to regulate RBP-binding (Cruz and Westhof, 2009). More specifically, the structure of an RNA molecule can make RBP recognition sequences inaccessible to a single-stranded RNA (ssRNA) binding protein, or promote binding by a double-stranded RNA (dsRNA) binding protein, or vice versa (Cruz and Westhof, 2009). Therefore, both the RNA sequence and its secondary structure are important *cis* regulators of RNA-protein interactions.

In this study, we utilized our protein interaction profile sequencing (PIP-seq) technique to simultaneously probe RNA secondary structure and RNA-protein interactions in the nuclei of *Arabidopsis* root hair and nonhair cells. This analysis revealed cell type-specific secondary structure and RBP binding patterns, some of which influence root epidermal cell development. Additionally, these protein-bound sequences were used to identify two RBPs, SERRATE and GLYCINE-RICH PROTEIN 8 (GRP8), that both regulate proper hair cell development. Together, these data elucidate novel post-transcriptional regulators of the plant root epidermal cell fate decision and development.

3.2 RESULTS

3.2.1 PIP-seq identifies thousands of cell type-specific protein-bound sites

To identify the differences in the nuclear RNA-protein interaction and RNA secondary structure landscapes of root hair and nonhair cells, we used the isolation of nuclei tagged in specific cell types (INTACT) method (Deal and Henikoff, 2010; Wang and Deal, 2015) to obtain highly pure nuclear samples (**Figure 3.1**). This technique utilizes cell type-specific promoters to drive expression of a fusion protein that targets a biotin ligase receptor peptide to the nuclear envelope. Therefore, by using plants that express this fusion protein under the control of the *ADF8* or *GL2* promoters we were able to specifically purify nuclei from root hair and nonhair cells, respectively (**Figure 3.2**). In fact, we obtained highly pure nuclei from both cell types that were completely devoid of the cytoplasmic and rough endoplasmic reticulum markers EIF1A, ALDOLASE, and CNX1 (**Figure 3.1**). These highly pure nuclei were then used for subsequent PIP-seq analyses.

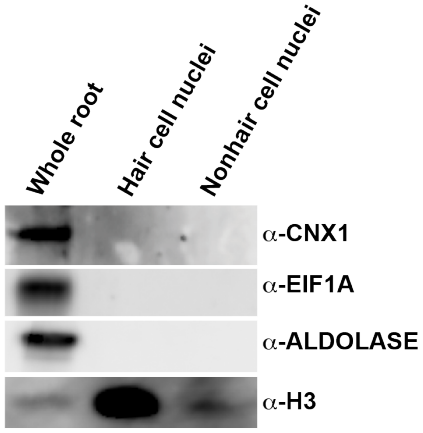


Figure 3.1: PIP-seq was performed on highly pure nuclei
Western blot analysis of whole root tissue, hair cell nuclei, and nonhair cell nuclei. H3 is used as a positive control for purified nuclei, while CNX1, EIF1A, and ALDOLASE are markers of the membrane of the endoplasmic reticulum and the cytoplasm. The non-nuclear protein markers are absent from the nuclear samples.

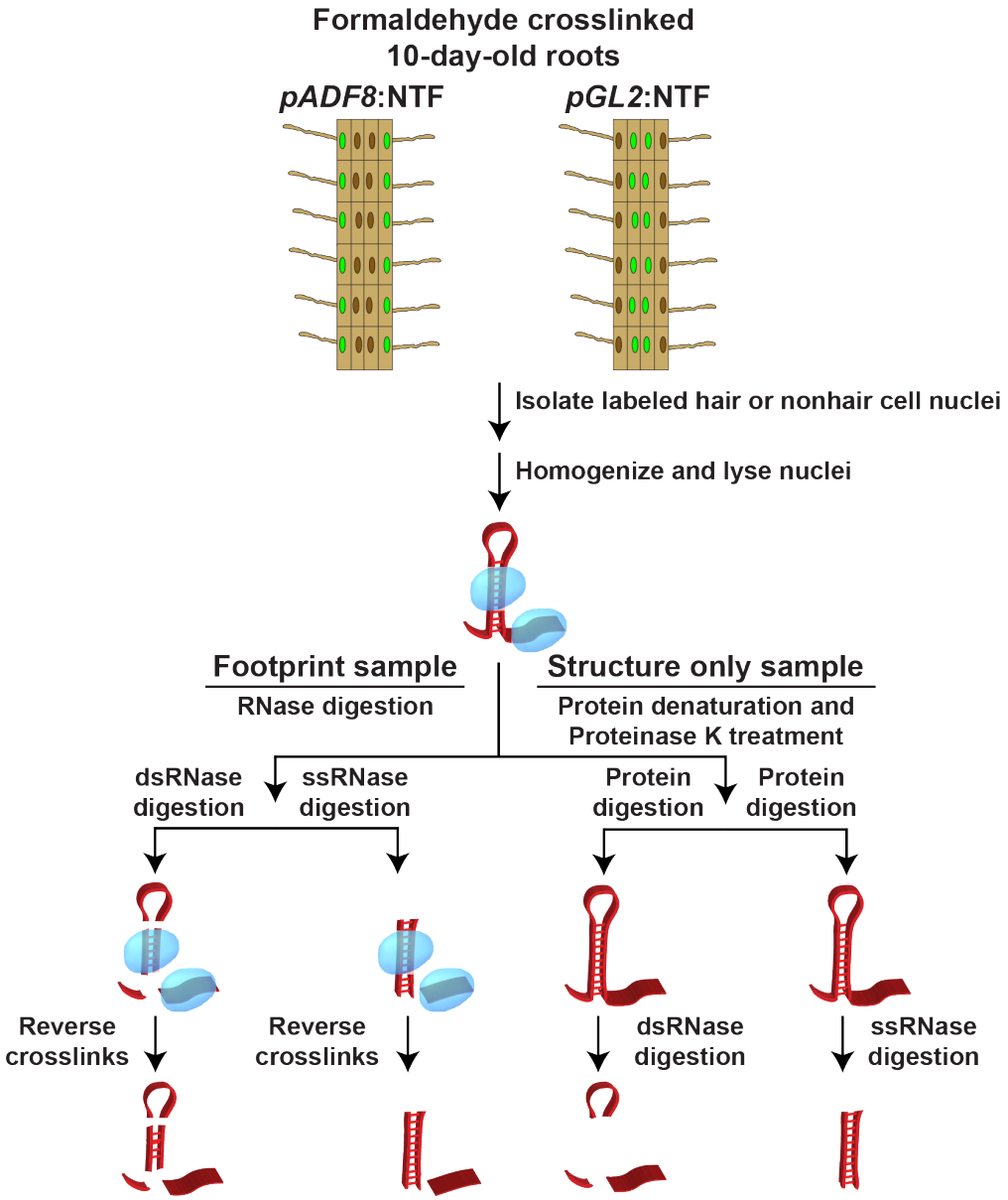


Figure 3.2: Nuclear PIP-seq identifies cell type-specific RNA-protein interactions. The PIP-seq approach in the nucleus of *Arabidopsis* root hair and nonhair cells. Fully differentiated root epidermal cells were excised from 10-day-old *Arabidopsis* plants and crosslinked with a 1% formaldehyde solution. The nuclei of either root hair or nonhair cells (green circles) were then isolated via the INTACT technique. Nuclei were lysed and separated into footprinting and structure only samples. Four total sequencing libraries were then prepared for each replicate experiment as previously described (Gosai et al., 2015).

PIP-seq allows global identification of RNA-protein interaction sites as well as RNA secondary structure (**Figure 3.2**) (Gosai et al., 2015; Silverman et al., 2014). We used ~2 million highly pure nuclei (**Figure 3.1**) for each of two PIP-seq replicates per cell type. These nuclei were

lysed, then divided into footprinting and structure only samples (four total libraries per replicate) (**Figure 3.2**). To globally identify RBP-bound RNA sequences, footprinting samples were directly treated with an RNase specific to either ssRNA or dsRNA (ssRNase or dsRNase, respectively), followed by protein denaturation and sequencing library preparation. In contrast, the structure only samples first had proteins denatured in SDS and degraded with Proteinase K prior to RNase digestion. Denaturation of proteins before RNase treatment makes sequences that were RBP-bound in the footprinting sample accessible to RNases in these reactions. Thus, sequences that are enriched in footprinting relative to structure only samples are identified as protein protected sites (PPSs) (Gosai et al., 2015; Silverman et al., 2014) (**Figure 3.2**). Additionally, using the structure only libraries allowed us to determine the native (protein-bound) RNA base-pairing probabilities for the nuclear transcriptomes of *Arabidopsis* root hair or nonhair cells, as previously described (Gosai et al., 2015; Li et al., 2012b).

The resulting PIP-seq libraries were sequenced and provided ~25-35 million raw reads per library. To determine reproducibility, we used a principle component analysis of read coverage in 100 nucleotide (nt) bins. This revealed that biological replicates of each library from the distinct cell types cluster together (**Figure 3.3A**), indicating the high quality and reproducibility of our root hair and nonhair nuclear PIP-seq libraries.

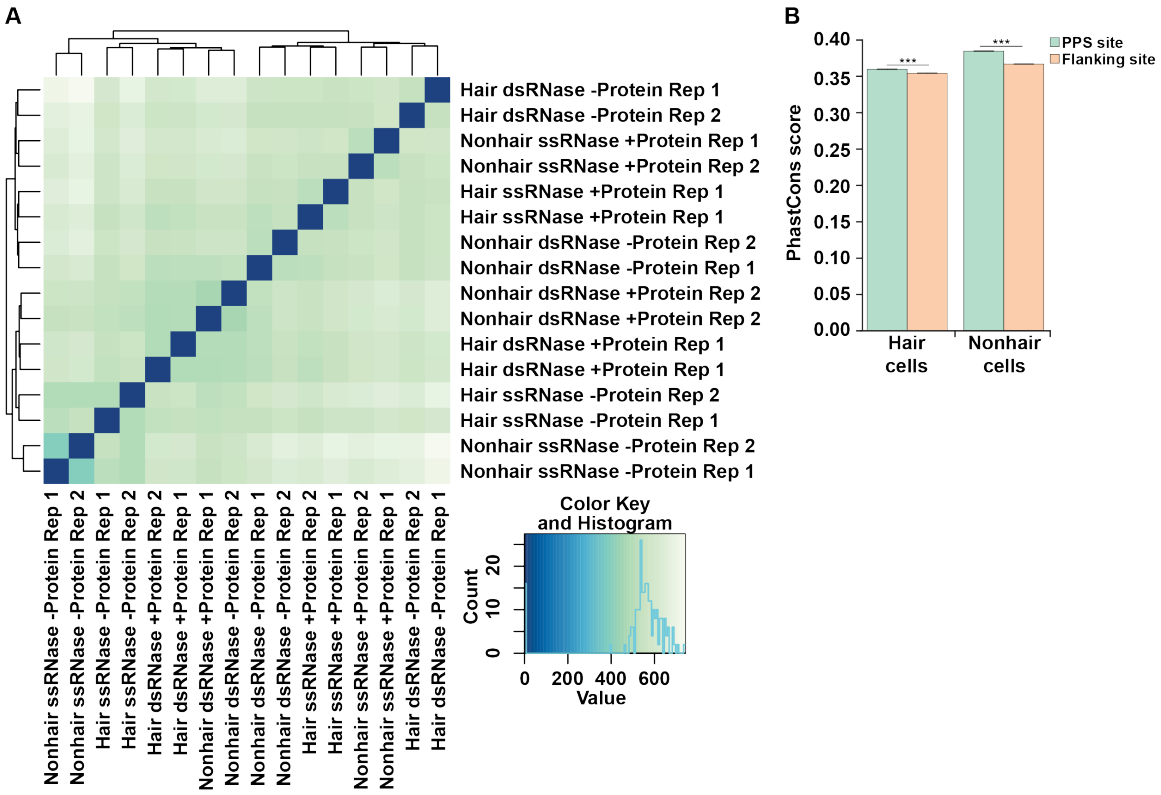


Figure 3.3: High quality PIP-seq was performed on highly pure nuclei
(A) Clustering analysis of all 16 PIP-seq libraries. The TAIR10 genome was divided into 100 nt bins and mapped reads were counted for each bin. The libraries were then clustered, with the most similar libraries (the biological replicates) clustering together. (B) Comparison of average PhastCons scores between PPSs (green bars) and equal sized flanking regions (orange bars). *** denotes *p* value < $1x10^{-10}$, Fisher's t-test.

To identify PPSs, we used a Poisson distribution model to identify enriched regions in the footprinting compared to the structure only libraries at a false discovery rate (FDR) of 5%, as previously described (Gosai et al., 2015). We identified a total of 34,442 and 44,315 PPSs in root hair and nonhair cell nuclei, respectively. To estimate the functional relevance of these nuclear PPSs from both root cell types, we compared flowering plant PhastCons conservation scores (Li et al., 2012b) for PPSs and equal-sized flanking regions. We found that PPS sequences were significantly (*p* values < 1.2×10^{-71}; KS-test) more evolutionarily conserved than flanking regions in both hair and nonhair cells (**Figure 3.3B**), indicating that there is evolutionary pressure to constrain these sites, likely due to their ability to interact with RBPs (Gosai et al., 2015).

Additionally, we observed a high overlap of PPSs between biological replicates. Whereas

CLIP-seq experiments will often find <35% of protein-bound sites shared between biological replicates (Lebedeva et al., 2011), we observed ~72% of dsRNase identified PPSs, and ~57% of ssRNase identified PPSs found in our first replicate are shared between both biological replicates (**Figures 3.4A-B**), with ~55-64% of hair cell and 27-36% of nonhair cell PPSs being identified by both ssRNase and dsRNase treatments (**Figures 3.4C-D**).

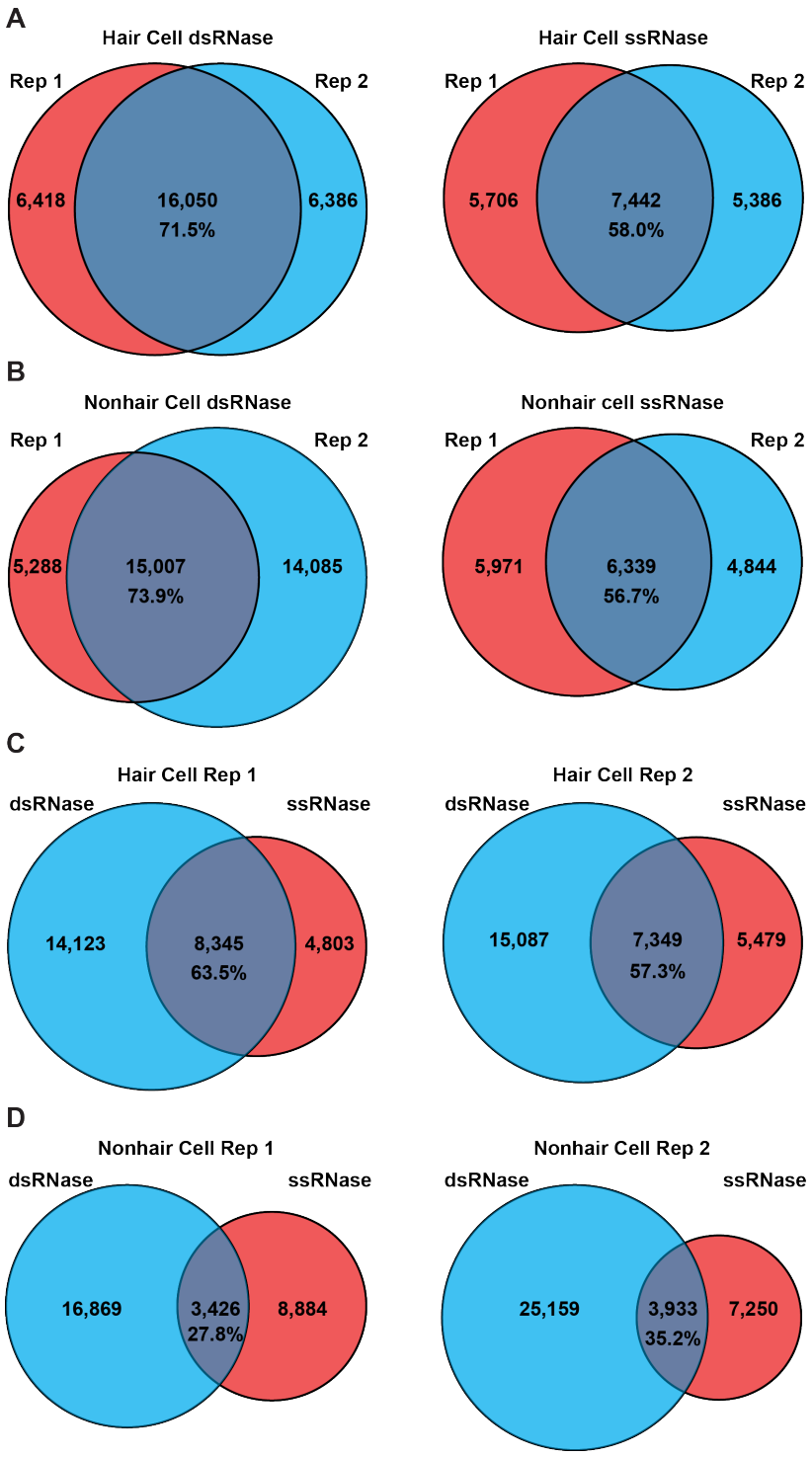


Figure 3.4: PPSs identified by PIP-seq are highly reproducible
(A-B) The overlap in PPSs present in both dsRNase- and ssRNase-treated libraries in hair cells (A) and nonhair cells (B). (C-D) The overlap in PPSs between biological replicates of root hair cell (C) and root nonhair cell (D) nuclei. Overlap is defined by at least one nucleotide overlapping between PPSs.

When comparing total identified PPSs found in hair cells, we observed 25,069 (72.8%) PPSs are also present in nonhair cells (**Figure 3.5A**). We next confirmed that these are true differences in protein occupancy at cell type-specific PPSs, rather than a representation of differentially expressed mRNAs. To do this, we analyzed PPSs present only in mRNAs expressed in both hair and nonhair cells. We found that the PPSs from both hair and nonhair cells within this subset of transcripts displayed an overlap of 73.4% (**Figure 3.5B**). Interestingly, we found 16,460 (72.4%) of dsRNase identified hair cell PPSs are common to both cell types (**Figure 3.5C**), whereas only 4,323 (34.4%) of ssRNase identified PPSs are common (**Figure 3.5D**), with the remaining 4,286 shared PPSs being identified in the dsRNase-treated sample of one cell type and the ssRNase-treated sample of the other cell type. Therefore, these data reveal that many cell type-specific protein-bound events are present in ssRNase-accessible regions.

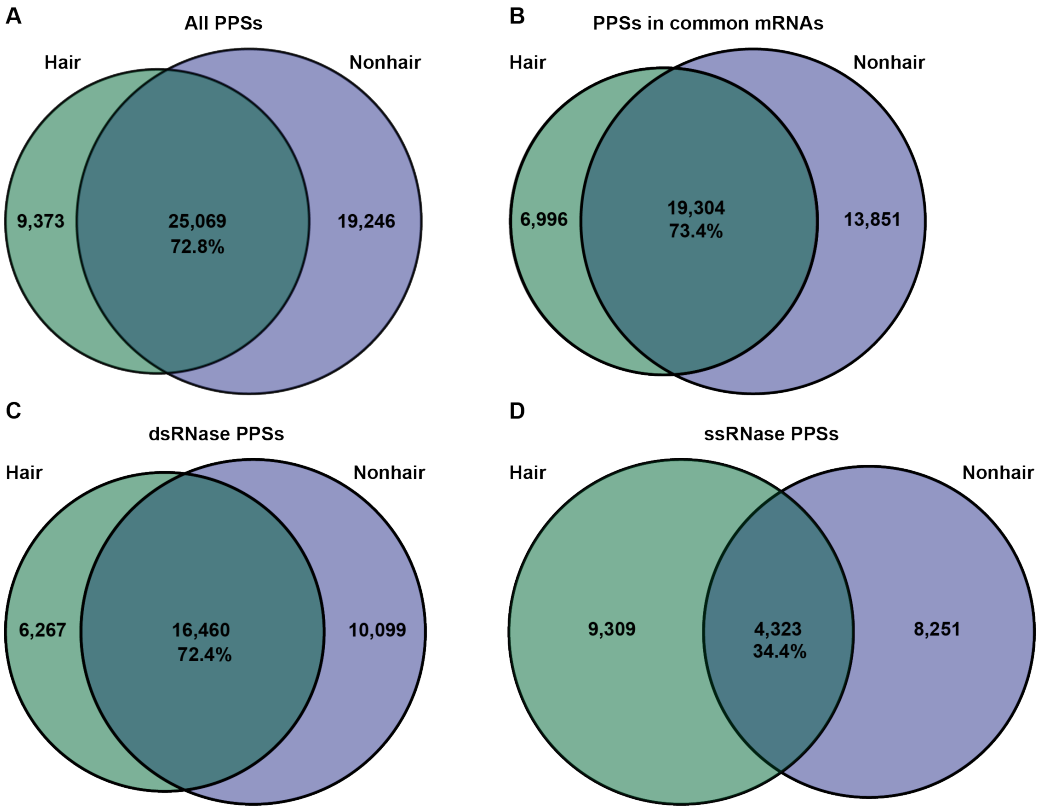


Figure 3.5: Cell type-specific PPSs are mostly identified by ssRNase treatment and present in transcripts expressed in both cell types.
(A) Overlap between protein protected sites (PPSs) identified in hair (green) or nonhair (purple) cell nuclei. (B) The overlap in PPSs only found in transcripts that are expressed in both root hair

and nonhair cells with more than 50 reads of depth. (C-D) The overlap in PPSs identified in dsRNase- (C) and ssRNase-treated (D) samples in the nuclei of root hair and nonhair cells. The intersection indicates PPSs identified in both cell types that overlap by at least a single nucleotide.

A classification of hair and nonhair cell PPSs revealed that >90% of these sites are localized to mRNAs, with the largest fractions occupying the coding sequence (CDS; ~55%) and introns (~25%) in both cell types (**Figure 3.6A**). We then tested the enrichment of PPSs in specific nuclear mRNA regions (e.g., CDS, introns, etc.) by comparing the number of PPS occupied nucleotides to the number of bases annotated as each feature in the TAIR10 *Arabidopsis* genome. We found that PPSs identified in both cell types were enriched in CDSs, while underrepresented in both untranslated regions (UTRs). Furthermore, introns showed a slight enrichment for PPSs in hair cells, but an underrepresentation in nonhair cells (**Figure 3.6B**). These findings are consistent with our previous results using nuclei isolated from whole seedlings (Gosai et al., 2015), both of which indicate that CDSs are highly RBP-bound in plant nuclei.

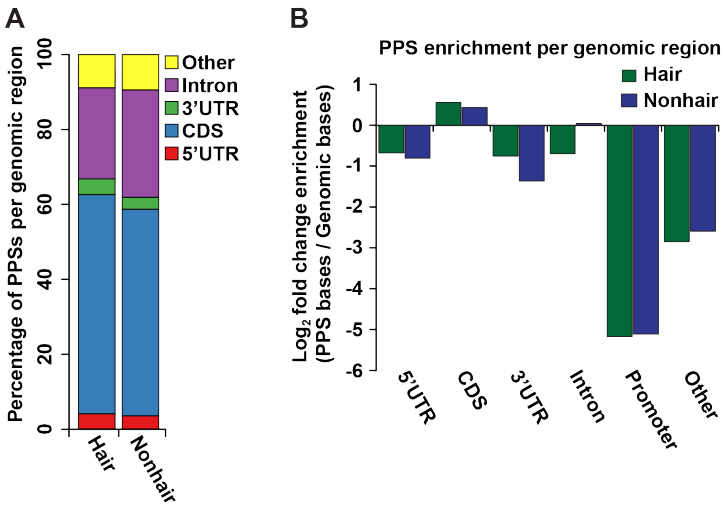
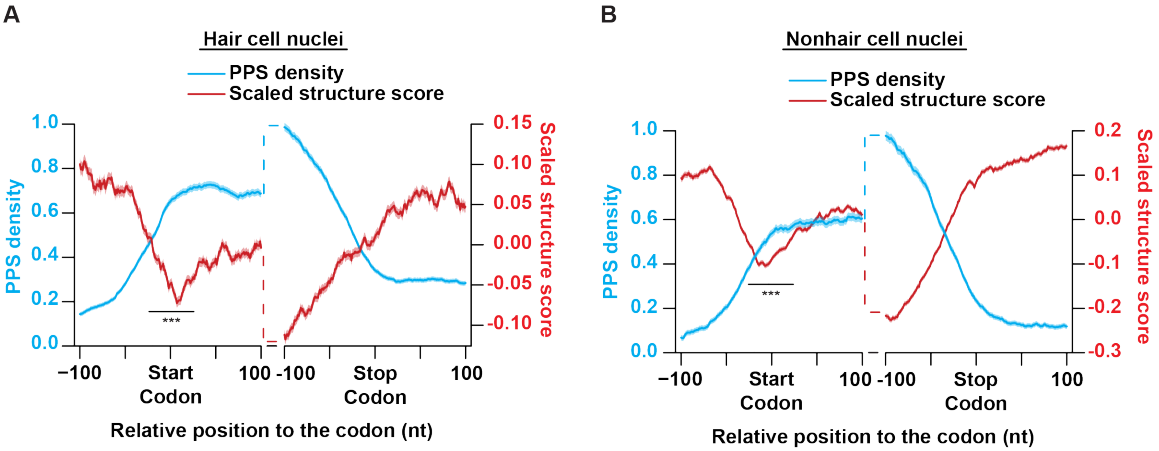


Figure 3.6: PPSs are primarily present in mRNAs, and enriched in the coding sequence
(A) Absolute distribution of PPSs throughout regions of mRNA transcripts. (B) Genomic enrichment of PPS density, measured as $\log_2$ enrichment of the fraction of PPS base coverage normalized to the fraction of annotated genomic bases of indicated mRNA regions for hair (green bars) and nonhair (purple bars) cells.

3.2.2 Hair and nonhair cells have distinct RNA-protein interaction and RNA secondary structure profiles in shared mRNAs and lncRNAs

To interrogate the landscape of RBP binding and RNA secondary structure in specific regions of nuclear mRNAs expressed in both hair and nonhair cells, we first calculated their structure scores and PPS densities. The structure score is a generalized log ratio of ds- to ssRNA-seq reads at each nucleotide position. These raw scores are then scaled by generating Z-scores (Berkowitz et al., 2016), with positive and negative scores indicating high likelihood of ds- and ssRNA, respectively. To examine the relationship between PPS density and structure score, we focused on the 100 nt up- and downstream of the start and stop codons of nuclear mRNAs expressed in both cell types. From this analysis, we observed the highest PPS density in the CDS with decreased occupancy within the 5' and 3' UTRs (**Figures 3.7A-B**), consistent with the overall PPS localization and enrichment analysis (**Figure 3.6**).



Region	*Spearman's rho*
All scores	-0.31
5' UTR	-0.96
5' CDS	-0.05
3' CDS	-0.99
3' UTR	-0.29

Region	*Spearman's rho*
All scores	-0.33
5' UTR	-0.90
5' CDS	0.91
3' CDS	-0.99
3' UTR	-0.90

Figure 3.7: RNA secondary structure and RNA-protein interactions are anti-correlated in the nuclei of both cell types
(A-B) PPS density (blue line) and scaled structure score (red line) profiles for nuclear mRNAs at each nucleotide +/- 100 nt from the annotated start or stop codons in hair (A) or nonhair (B) cell nuclei. The tables represent the Spearman's rho correlations between the PPS density and scaled structure scores across the graphed windows up- and downstream of the start codon, stop codon, or across all detectable mRNA transcripts. Shading around the solid lines indicates

standard error of the mean (SEM) across all detectable transcripts. *** indicates *p* value < 0.001, Wilcoxon test in all panels.

In contrast to RBP occupancy, we found that secondary structure was higher in both UTRs compared to the CDS within the regions analyzed in both cell types. Additionally, we observed a significant (*p* values < 6.6×10^{-13}; Wilcoxon test) dip in secondary structure directly over start codons, as well as upstream of the stop codon (**Figures 3.7A-B**), two characteristics which have been observed in numerous studies of RNA secondary structure across various organisms (Ding et al., 2014; Gosai et al., 2015; Li et al., 2012a, 2012b). Additionally, all of these results are consistent with the patterns observed previously for nuclear mRNA secondary structure from whole seedling nuclei (Gosai et al., 2015). Thus, these structural characteristics across the UTRs and CDS seem to be a consistent feature of the *Arabidopsis* nuclear mRNA transcriptome.

Consistent with our study of whole seedling nuclei, our combined analyses of RBP binding and RNA secondary structure revealed that these features are anti-correlated across nuclear mRNAs (Spearman's rho ≤ -0.31; *p* value < 2.2×10^{-16}; asymptotic *t* approximation) in both root epidermal cell types. In addition to this transcriptome-wide pattern for both cell types, we found even stronger anti-correlations (Spearman's rho ≤ -0.90; *p* value < 2.2×10^{-16}; asymptotic *t* approximation) between protein binding and RNA folding within the last 100 nt of 5' UTRs and CDSs of nuclear mRNAs expressed in both hair and nonhair mRNAs. Interestingly, we observed a discrepancy within the first 100 nt of mRNA 3' UTRs from root hair and nonhair cells. Specifically, we found a strong negative correlation (Spearman's rho ≤ -0.99; *p* value < 2.2×10^{-16}; asymptotic *t* approximation) between protein binding and structure in nonhair cell nuclei, with a much more mild correlation (Spearman's rho ≤ -0.29; *p* value < 0.0036; asymptotic *t* approximation) in hair cell nuclei. This distinct pattern indicates that there may be differential protein binding in the 3' UTRs of these two cell types.

Conversely, the RBP binding and RNA secondary structure of the first 100 nt of the CDS did not exhibit an anti-correlation. We found no significant correlation in hair cells, as well as a

significant positive correlation (Spearman's rho > 0.91; *p* value < 2.2 x 10^{-16}; asymptotic *t* approximation) in nonhair cells. These observations are striking as they oppose the anti-correlation that we found in this same region when profiling mixed nuclei from whole seedlings (Gosai et al., 2015). Taken together, these observations reveal a cell type-specific interplay between RNA folding and RBP binding near the start codon of nuclear mRNAs. Given that these results are from highly pure nuclear samples (**Figure 3.1**), the PPSs cannot simply indicate ribosome binding, and are likely caused by cell type-specific RBP interactions. Identifying and characterizing these proteins will be the focus of future inquiry.

We next aimed to directly compare the RNA secondary structure patterns in the nuclei of these two cell types by using these scaled structure scores. We found that RNA secondary structure is similar in both cell types within the 200 nt window flanking the start codon (**Figure 3.8A**). Conversely, there are significant (*p* values < 3.1x10^{-4}; Wilcoxon test) differences in RNA secondary structure within the 100 nt windows up- and downstream of the stop codons of mRNAs found in both hair and nonhair cell nuclei. Specifically, we found significantly higher RNA secondary structure in these mRNAs within the last 100 nt of their CDSs in hair compared to nonhair cells, while the opposite pattern was observed for the first 100 nt of their 3' UTRs (*p* values < 1.7x10^{-5} and 2.2x10^{-16}, respectively; Wilcoxon test) (**Figure 3.8A**). These differences in secondary structure around the stop codon could provide an intriguing mechanism for regulating RBP binding within these specific transcript regions. Therefore, we also directly compared the density of hair and nonhair cell-specific as well as common PPSs in the 200 nt regions surrounding the start and stop codons of mRNAs expressed in both hair and nonhair cells (**Figure 3.B**). Although overall RBP binding had a similar profile across mRNAs from both cell types (**Figures 3.7A-B**), there is a significant (*p* value < 3.4x10^{-15}; Wilcoxon test) increase in hair cell-specific RBP binding events within the first 100 nt of the 3' UTRs of mRNAs expressed in both cell types (**Figure 3.8B**). These findings are consistent with the significantly (*p* value < 2.2x10^{-16}; Wilcoxon test) decreased RNA secondary structure also observed in this transcript region in hair compared to nonhair cells (**Figure 3.8A**), given that these features are generally

anti-correlated with one another (**Figures 3.7A-B**). Thus, this nuclear PIP-seq analysis reveals cell type-specific differences in both RNA secondary structure and RBP binding profiles between hair and nonhair cells. In total, our findings suggest that cell type-specific RNA folding and RBP binding in protein-coding mRNAs is a likely mechanism for differential regulation of the root hair and nonhair cell transcriptomes, and the resulting cell fate decisions.

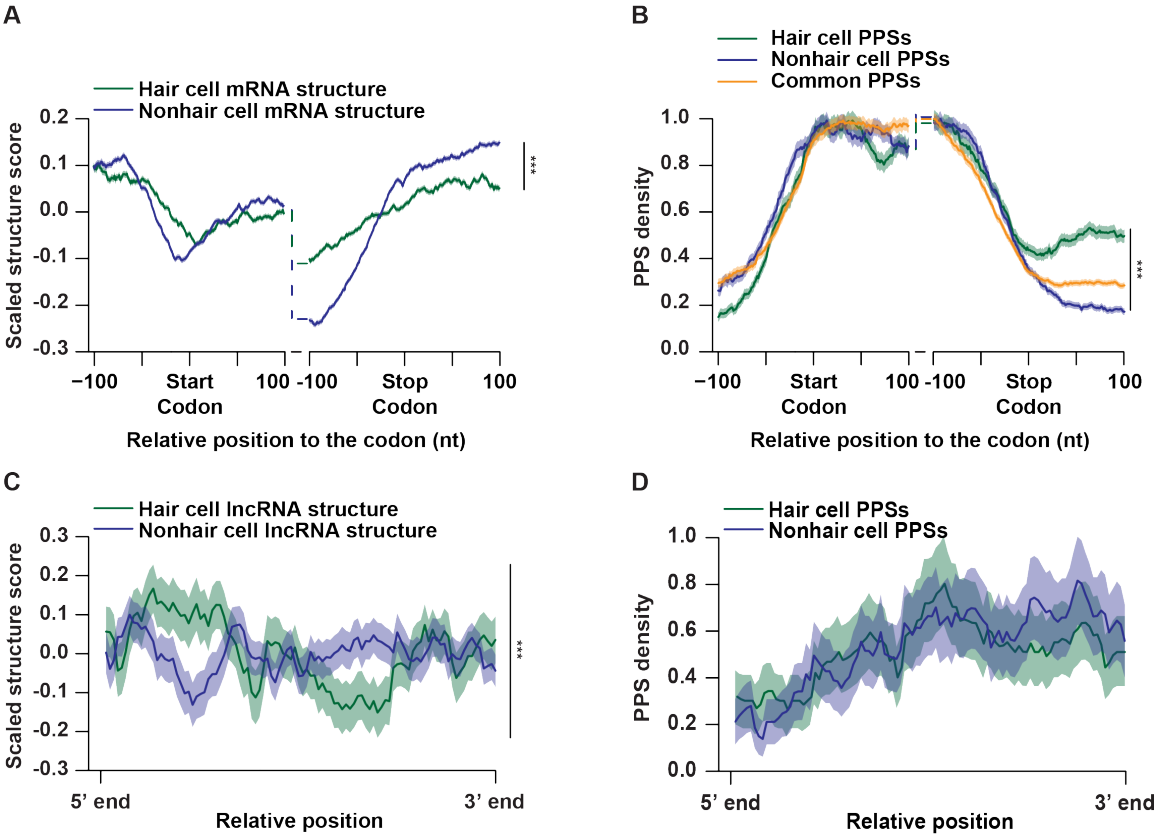


Figure 3.8: Hair and nonhair cells have distinct RNA-protein interaction and RNA secondary structure profiles.
(A-B) Scaled structure score (A) or PPS density (B) profiles at each nucleotide +/- 100 nt from the annotated start or stop codons in nuclear mRNAs expressed in both hair (green line) and nonhair (purple line) cells. PPSs are divided into those that are detected in hair cells (green line), nonhair cells (purple line), or common to both cell types (orange line). (C-D) Scaled structure score (C) or PPS density (D) across binned unspliced lncRNAs expressed in root hair (green) or nonhair (purple) cell nuclei. Shading around the solid lines indicates standard error of the mean (SEM) across all detectable transcripts. *** indicates *p* value < 0.001, Wilcoxon test in all panels.

In addition to mRNAs, we examined both RNA secondary structure and RNA-protein interactions in long noncoding RNAs (lncRNAs) that are found in the nucleus. Using a comprehensive list of *Arabidopsis* lncRNAs (Liu et al., 2012), we first analyzed the secondary

structure of these transcripts in root hair and nonhair cell nuclei (**Figure 3.8C**).Taking the entire length of the unspliced annotated lncRNAs, we divided each transcript into 100 equally sized bins. Graphing the average scaled structure score of each bin, we found significant (*p* value < $2.2x10^{-16}$; Wilcoxon test) differences between the structure profiles of the lncRNAs found in both root hair and nonhair cells. Specifically, these lncRNAs in root hair cell nuclei exhibited increased structure at the 5' end of the transcript, while being less structured near the 3' end than these lncRNAs in nonhair cell nuclei (**Figure 3.8C**). As the structural profiles differ dramatically, we next examined PPS binding across lncRNAs. This analysis revealed that a vast majority (>82%) of lncRNA mapping PPSs in hair cells are shared with nonhair cell nuclei. Unsurprisingly, when graphing the PPS density across all lncRNAs identified in root hair or nonhair cells, these profiles were not significantly different (*p* value > 0.05; Wilcoxon test) (**Figure 3.8D**). Therefore, like mRNAs, lncRNAs exhibit cell type-specific secondary structure. However, unlike mRNAs, these differences do not result in a significant difference in RBP binding across these transcripts. Although these transcripts are bound by similar numbers of proteins in each cell type, this difference in secondary structure likely indicates that differing cohorts of proteins are binding lncRNAs in hair and nonhair cells.

3.2.3 SERRATE regulates root hair length and hair cell fate in a microRNA-independent and a microRNA-dependent manner, respectively

To determine whether cell type-specific RBP binding regulates the root hair and nonhair cell fate decision, we identified RBPs that function in a cell type-specific manner. It is worth noting that all PPSs correspond to protein-bound sites in fully differentiated cells, rather than non-terminally differentiated precursor cells. Therefore, only protein binding sites established in these precursor cells, and maintained after differentiation will lead to the identification of RBPs necessary for cell fate decision. To identify those PPSs potentially necessary for differentiation, we subsetted all identified PPSs into those that are hair or nonhair cell-specific as well as those common to both cell types (**Figure 3.5A**). Taking these three subsets of RBP-bound sequences,

we used the motif finding algorithm MEME (Bailey et al., 2009) to identify enriched protein-bound sequences. We identified a combined 54 significantly (*E* values < 0.01) enriched motifs across these three subsets.

To identify the specific RBPs that interact with a subset of these motifs, we performed RNA affinity chromatography followed by mass spectrometry analysis. In this technique, we covalently attached a synthetic RNA motif or a scrambled sequence control to agarose beads. We then incubated these RNA baits, as well as a bead-only control, with whole root lysate, and stringently washed away any weakly bound proteins. The specifically bound proteins were identified via mass spectrometry. Using this approach, we identified 58 annotated RBPs that are at least 4-fold enriched for interaction with at least one of the twelve tested sequence motifs, as compared to the scrambled sequence and bead-only negative controls (**Figure 3.9**). One motif of particular interest, a GGN repeat motif that was enriched in PPSs common to both root hair and nonhair cell nuclei, was found to interact with the RBP SERRATE (SE) (AT2G27100) (**Figure 3.10A**). SE is known to function in conjunction with ABA HYPERSENSITIVE 1/CAP-BINDING PROTEIN 80 (ABH1/CPB80, AT2G13540) and HYPONASTIC LEAVES 1 (HYL1, AT1G09700) in microRNA (miRNA) biogenesis, where these three RBPs recruit DICER-LIKE 1 (DCL1, AT1G01040) to primary miRNA transcripts to allow their processing to mature miRNAs (Dong et al., 2008; Yang et al., 2006). Additionally, SE and ABH1/CBP80 regulate alternative splicing across the *Arabidopsis* transcriptome. This variety of functions indicated that SE was a reasonable candidate as a potential regulator of root hair cell fate. We first confirmed that SE is expressed at similar levels in both root hair and nonhair cell nuclei (**Figure 3.10B**), then performed RNA immunoprecipitation (RIP) followed by RT-qPCR to confirm that SE interacts with transcripts containing the GGN repeat motif *in vivo*. To do this, we incubated lysates from formaldehyde crosslinked roots with polyclonal α-SE, α-ABH1/CBP80, or the negative control rabbit IgG. We first confirmed pulldown of SE and ABH1/CBP80 by these antibodies (**Figures 3.10C-D**), then performed RT-qPCR for 13 GGN repeat containing mRNAs. We found that all 13 of the transcripts were significantly (all *p* values < 0.05; Welch's t-test) enriched >1.5-fold in the α-

SE compared to the IgG control RIP samples, as opposed to the *ACTIN2* negative control (**Figure 3.10E**). Furthermore, none of the 13 transcripts were enriched in α-ABH1/CBP80 compared to the IgG control RIP samples. Taken together, these findings indicate that SE interacts *in vivo* with GGN motif-containing mRNAs, while ABH1/CBP80 does not.

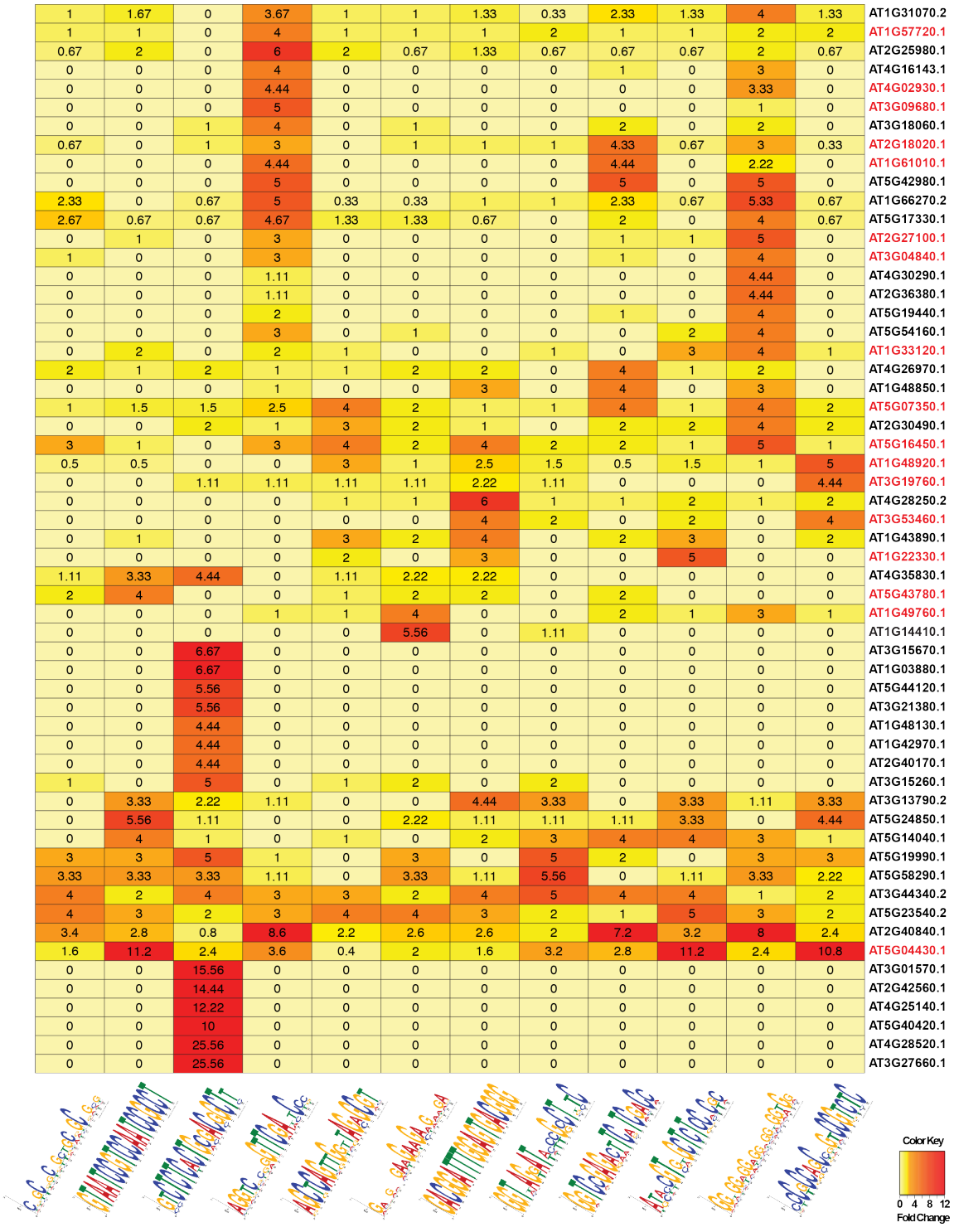


Figure 3.9: Numerous RBPs are identified via RNA-affinity chromatography
A heatmap showing enrichment of proteins in motif sample compared to the scramble and bead-only controls after RNA affinity chromatography. These are candidate RBPs that have more than 2-fold enrichment in at least one sample. Red text indicates an annotated RBP.

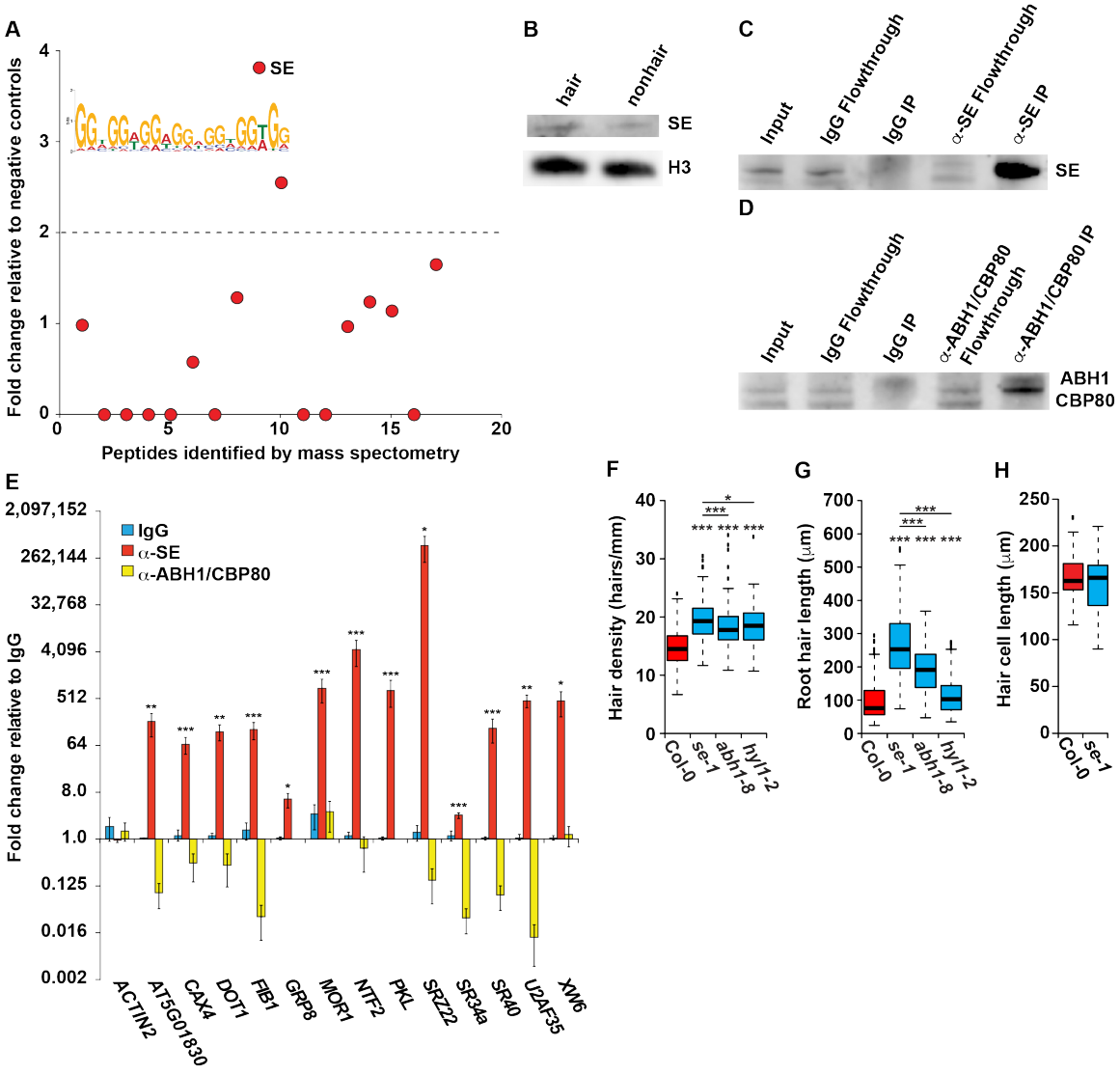


Figure 3.10: SERRATE regulates hair cell fate and hair length in a partially microRNA-independent manner

(A) RNA affinity chromatography followed by LC-MS was performed on whole root cell lysate using the MEME identified GGN repeat motif as bait. The number of peptide spectrum matches (PSMs) for each identified peptide was graphed as fold change over the average PSMs of scrambled RNA bait and no RNA controls. Peptides above the dotted line have a more than 2-fold change and correspond to candidate RBPs. SE is denoted as being highly bound by our analysis. (B) Western blot of nuclear lysate from root hair and nonhair cells examining SE expression using H3 as a loading control. There is no detectable difference in SE levels once normalizing to H3. (C-D) Western blot of immunoprecipitations performed with α-SE (C), α-ABH1/CBP80 (D), or rabbit IgG (negative control) performed on whole seedlings. (E) RIP-qPCR was performed on whole root lysate using rabbit α-IgG (blue bars), α-SE (red bars), or α-ABH1/CBP80 (yellow bars) antibodies, graphed as fold change relative to the IgG negative control pull down, n = 4. *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, Welch's t-test. Error bars indicate SEM. (F-H) Root hair cell density (hairs/mm) (F), root hair length (μm) (G), and hair cell length (H) of Col-0, *se-1, abh1-8,* and *hyl1-2* mutant plants. For

analysis of root hair length n=400, for root hair density n > 135, and for root hair cell length n > 30 *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, while N.S. denotes *p* value > 0.05, Wilcoxon test.

After validating *in vivo* GGN motif-containing mRNA binding by SE, we next determined whether this protein regulates root hair cell fate and development. To do this, we measured the root hair cell density (hairs/mm) and root hair length in 8-day-old wildtype Col-0 (hereafter WT) and *SE* hypomorphic (*se-1*) seedlings (Clarke et al., 1999; Serrano-Cartagena et al., 1999). From this analysis, we found that *se-1* mutant seedlings had significantly (*p* values < $2.2x10^{-16}$; Wilcoxon test) more root hair cells that are significantly (*p* values < $2.2x10^{-16}$; Wilcoxon test) longer as compared to WT (**Figures 3.10F-G**), indicating that SE functions in both promoting root nonhair cell fate and terminating root hair extension. The difference in hair cell density on *se-1* plants could be caused by either promotion of hair cell fate, resulting in ectopic hair cells, or by decreased epidermal cell size, packing hair cells closer together. Therefore, we measured the size of hair cell bodies and found that there is no significant (*p* value > 0.05; Wilcoxon test) difference in their size in *se-1* compared to WT roots (**Figure 3.10H**). Combined, these findings demonstrate that SE functions both in precursor epidermal cells to promote nonhair cell fate, as well as in differentiated hair cells to terminate hair growth. This variety of functions is unsurprising as this RBP binds (**Figure 3.10E**) and post-transcriptionally regulates many different transcripts (Clarke et al., 1999; Raczynska et al., 2014).

We next tested whether SE influences the cell fate decision by functioning in the same pathway as the canonical transcription factors CAPRICE and WEREWOLF. Taking *CAPRICE* null (*cpc-1*) and wildtype WS ecotype *Arabidopsis* as a control, we measured levels of *SE* and *GL2* in plant roots, while 5S rRNA was used as a control. Given that *cpc-1* plants are enriched in nonhair cells, we observed the expected significant (*p* value < 0.05; Welch's t-test) increase in abundance of the nonhair cell-specific *GL2* marker gene. Conversely, no significant difference in *SE* level was observed between wildtype WS and *cpc-1* plants (**Figure 3.11A**), which was consistent with our western blot results (**Figure 3.10B**). Additionally, we measured the abundance of several transcription factors known to function in the CAPRICE/WEREWOLF pathway in the roots of *se-1*

plants, and also observed no significant (p value > 0.05; Welch's t-test) difference in their levels compared to wildtype plants (**Figure 3.11B**). Together, these data suggest that SE functions independently of the canonical CAPRICE/WEREWOLF pathway in determining root hair cell fate.

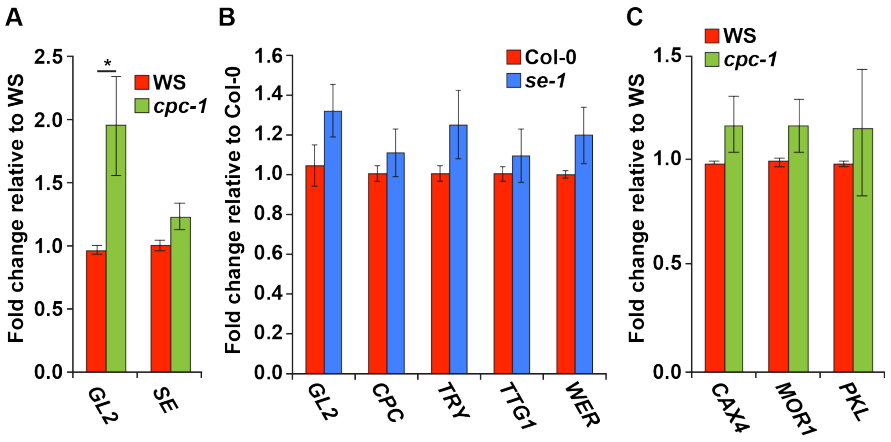


Figure 3.11: SE influences root hair cell fate independently of the CAPRICE/WEREWOLF transcription factor network.
(A) RT-qPCR using RNA from roots of WS (red) or *cpc-1* (light green) plants measuring *GL2* (positive control) and *SE* transcript levels. The results show no significant difference in *SE* abundance between genotypes. * indicates p value < 0.05; Welch's t-test. (B) RT-qPCR using RNA from root tissue of Col-0 (red) and *se-1* (blue) plants measuring components of the CAPRICE/WEREWOLF transcription factor network. The results show no significant difference in the abundance of the tested transcripts between the two genotypes. (C) RT-qPCR based measurements of *CAX4, MOR1,* and *PKL* levels in the roots of WS (red) and *cpc-1* (light green) plants showing no significant difference in the abundance of these transcripts between genotypes. * denotes p value < 0.05, Welch's t-test. Error bars indicate SEM.

As SE functions in both microRNA biogenesis and alternative splicing, our next goal was to differentiate the effect of these two regulatory mechanisms on its function in root hair cell development. To do this, we assayed for root hair length and density phenotypes in null mutants of *ABH1/CBP80* (*abh1-8*) and hypomorphic mutants of *HYL1* (*hyl1-2*), both of which are known to function in conjunction with SE during plant miRNA biogenesis. We measured root hair density for 8-day-old WT, *abh1-8*, and *hyl1-2* seedlings and found significant (p values < 5.6×10^{-15}; Wilcoxon test) increases in the density of root hairs in both *abh1-8* and *hyl1-2* mutant compared to WT (**Figure 3.10F**). These increases were similar in magnitude to those seen in the *se-1* mutant seedlings, indicating that this root hair cell fate phenotype is miRNA biogenesis dependent. Additionally, we found the root hair lengths in *abh-1* and *hyl1-2* seedlings to be significantly (p

values < 3.7 x 10^{-9}; Wilcoxon test) longer than those of WT. However, they are also significantly (*p* values < 2.2x10^{-16}; Wilcoxon test) shorter than those observed for *se-1* seedlings (**Figure 3.10G**). This mild increase in hair length in *abh1-8* and *hyl1-2* mutant roots indicates that decreased miRNA biogenesis in *se-1* plants accounts for a portion of the root hair length phenotype. However, there are also important SE-specific regulatory functions that add to the increased hair length observed in *se-1* mutant seedlings. Taken together, these findings reveal that although the function of SE in the microRNA biogenesis pathway is required for regulating root hair cell fate, this protein also has specific effects on root hair length.

In order to better understand these SE-specific effects on hair length, we investigated the root phenotypes of mutants lacking one of several GGN motif-containing genes that we found were bound by SE (**Figure 3.10E**). Although none of these genes are known to function in root hair cell fate, three of them have known roles in root development. CATION EXCHANGER 4 (CAX4, AT5G01490) promotes both primary and lateral root growth in plants subjected to Cd^{2+} toxicity (Mei et al., 2009). MICROTUBULE ORGANIZATION 1 (MOR1, AT2G35630) regulates microtubule assembly, and when temperature sensitive *mor1-1* mutants are grown at the restrictive temperature there is an increase in primary root diameter (Whittington et al., 2001). Additionally, the chromatin-remodeling factor PICKLE (PKL, AT2G25170) is necessary for silencing embryonic genes and promoting lateral root development (Furuta et al., 2011; Ogas et al., 1999). Interestingly, when screening 8-day-old seedlings lacking any one of these proteins (*cax4-1, mor1-1,* and *pkl1-1*) we found significantly (all *p* values < 0.001; Wilcoxon test) aberrant root hair length as compared to WT (**Figures 3.12A-C**). Specifically, we observed that *cax4-1* and *pkl1-1* mutant seedlings had longer root hairs (**Figures 3.12A and 3.12C**), similar to *se-1*. Conversely, we found that *mor1-1* mutant seedlings grown at the restrictive temperature displayed shorter root hairs compared to WT (**Figure 3.12B**). Taken together, these data suggest that the increased root hair length observed for *se-1* plants is likely due to the additive effects of misregulation of numerous mRNA substrates. We next confirmed that *CAX4, MOR1,* and *PKL* transcripts are not regulated by the CAPRICE/WEREWOLF transcription factor network by

measuring their abundance in the roots of wildtype WS and *cpc-1* plants. From this analysis, we observed no significant (*p* value > 0.05; Welch's t-test) difference (**Figure 3.12D**) in the levels of these three in the absence of CAPRICE function.

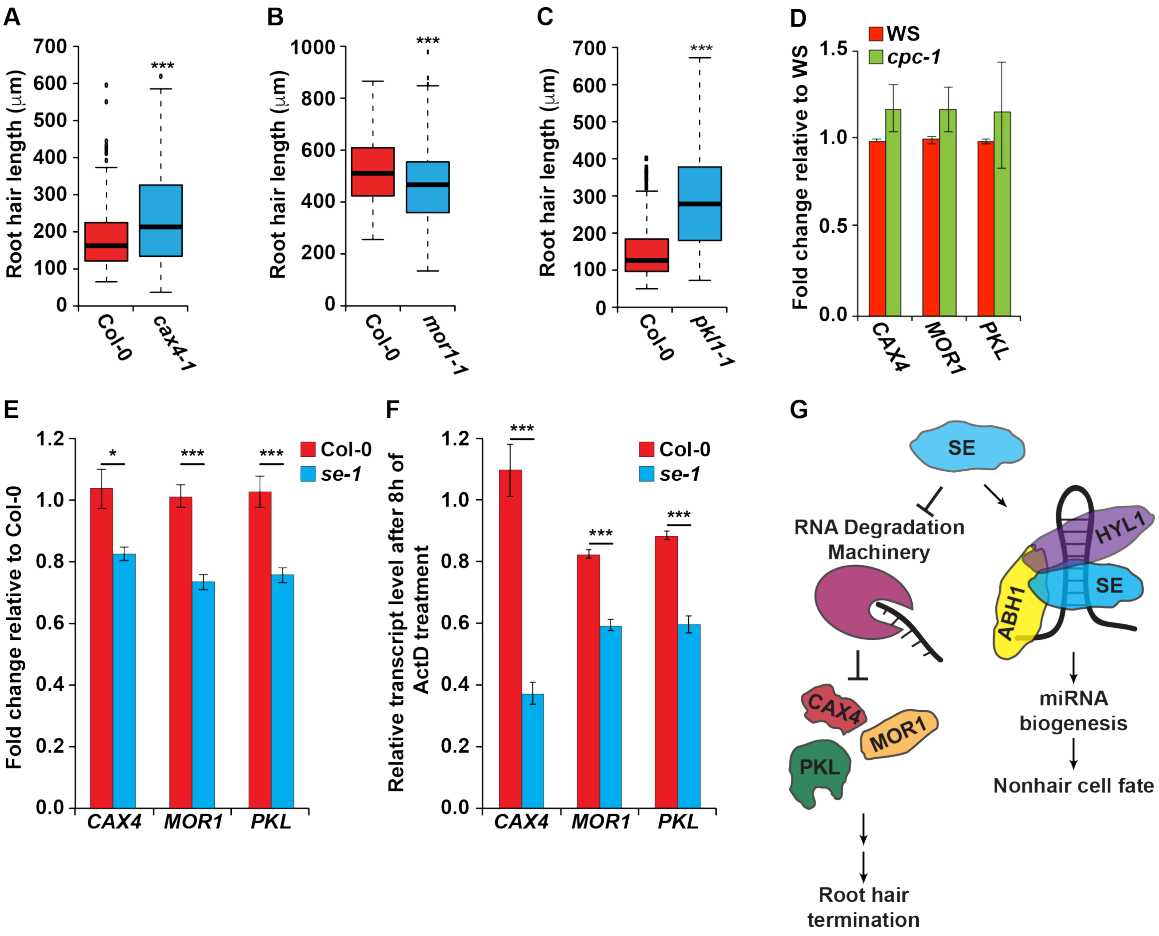


Figure 3.12: SE-bound GGN motif containing genes regulate root hair cell development
(A-C) Root hair length for null *cax4-1* (A), *mor1-1* (B), and *pkl1-1* (C) mutant plants as compared to wildtype Col-0. For root hair length analysis n=200. *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, Wilcoxon test. (D) RT-qPCR based measurements of *CAX4, MOR1,* and *PKL* levels in the roots of WS (red) and *cpc-1* (light green) plants showing no significant difference in the abundance of these transcripts between genotypes. (E) RT-qPCR of SE-bound genes in WT (red) and *se-1* (blue) roots, n = 6. (F) Roots from both WT (red) and *se-1* (blue) plants were subjected to Actinomycin D treatment for 8 hours to inhibit transcription, followed by RT-qPCR analysis of the mRNAs noted in the figure, n = 6. For (D)-(F) *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, Welch's t-test. Error bars indicate SEM. (G) A model of the role of SE in both the microRNA-independent promotion of root hair termination, as well as the microRNA-dependent promotion of the nonhair cell fate.

To test how SE affects the abundance of these RNAs, we measured their levels in the roots of WT and *se-1* plants. We found that all three transcripts are significantly (all *p* values <

0.05; Welch's t-test) decreased in *se-1* roots (**Figure 3.12E**), suggesting that SE stabilizes these transcripts. To further test this idea, we excised roots from 8-day-old WT and *se-1* plants and incubated them in liquid media containing the transcription inhibitor Actinomycin D (Act D). Specifically, we incubated these roots for 8 hours in this media, and then measured changes in transcript levels in the absence of transcription via RT-qPCR. This analysis revealed significantly (*p* value < 0.001; Welch's t-test) decreased transcript levels in *se-1* roots compared to WT, indicating decreased transcript stability in the absence of SE function (**Figure 3.12F**). In total, our results reveal that SE promotes the nonhair cell fate in a miRNA biogenesis-dependent manner, while also terminating root hair growth by stabilizing the mRNA transcripts of proteins involved in specifying hair length in plant roots (**Figure 3.12G**).

3.2.4 GRP8 regulates root hair cell fate independently of GRP7

As we had observed a dramatic difference in secondary structure and protein binding in the first 100 nt of 3' UTRs of root hair and nonhair cell mRNAs (**Figures 3.8A-B**), we next examined this region for enriched protein-bound motifs. To do this, we performed MEME on all hair or nonhair cell PPSs located in this area. While we did not observe any significant (E-value > 0.01) motifs in nonhair cell PPSs, we identified three significant (E-value < 0.01) motifs in hair cell PPSs (**Figure 3.13**). Although the two GA-rich motifs are similar to other motifs identified in nonhair cell PPSs and common PPSs, the TG-rich motif was only previously identified in hair cell-specific PPSs. As this motif is enriched in regions of differential protein binding, we next aimed to identify what proteins are able to bind to this sequence, and determine their role in root hair development.

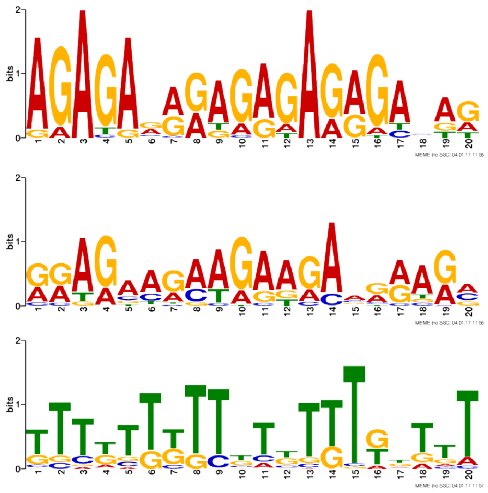


Figure 3.13: Enriched protein-bound motifs identified in hair but not nonhair cell PPSs localized to the first 100 nt of annotated mRNA 3' UTRs.
The first 100 nt of 3' UTRs annotated in TAIR10 were taken and intersected with all hair or nonhair cell PPSs. MEME was used to identify significantly (E-value < 0.01) enriched motifs in this region for hair cell PPSs, while no significant motifs were identified for nonhair cell PPSs. The figure shows the three hair cell-specific motifs found in this region, an area that shows a hair cell-specific decrease in RNA secondary structure and an increase in RBP binding.

We performed RNA affinity chromatography using this TG-rich motif, and found four annotated RBPs that were >10-fold enriched over our negative controls. In addition to RBP45A (AT4G54900), we found multiple members of the family of GLYCINE-RICH PROTEINs (GRPs), GRP2 (AT14G13850), GRP7 (AT2G21660), and GRP8 (AT4G39260) interacted with this sequence motif (**Figures 3.14A**). GRPs are nuclear localized hnRNP-like proteins (Streitner et al., 2012) that are required for numerous processes in plants, including responses to various biotic and abiotic stresses via their function in regulating both alternative splicing and microRNA biogenesis (Lewinski et al., 2016). Therefore, we were unsurprised when we observed that similar levels of GRP7 and GPR8 were present in the nuclei of both root hair and nonhair cells (**Figure 3.14B**). Using an antibody that recognizes both native GRP7 and GRP8, we performed RIP-qPCR to validate *in vivo* binding of GRP7/8 to TG-rich motif containing transcripts in formaldehyde-crosslinked whole root lysate. Given that both GRP7 and GRP8 are known to bind the *GRP8* transcript (Schöning et al., 2008), we used it as a positive control, and identified a significant (p value < 0.01; Welch's t-test) enrichment of this transcript in the α-GRP7/8 compared

to our rabbit IgG negative control pulldown (**Figure 3.14C**). Of the eight TG-rich motif containing mRNAs tested, we found six genes to be significantly (all *p* values < 0.05; Welch's t-test) enriched in the α-GRP7/8 compared to the IgG negative control pulldown (**Figure 3.14C**). These data reveal either GRP7, GRP8, or both proteins bind to TG-rich motif-containing transcripts *in vivo*.

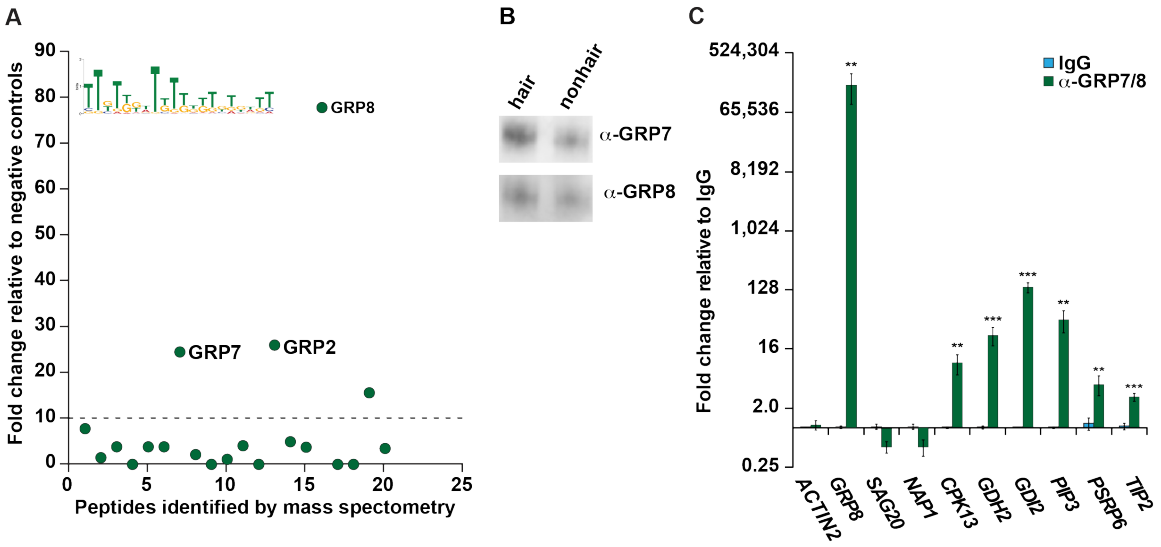


Figure 3.14: GRP7 and/or GPR8 bind the TG-rich motif *in vivo*
(A) RNA affinity chromatography followed by LC-MS was performed on whole root cell lysate using the MEME identified TG-rich motif as bait. Peptides above the dotted line have a more than 10-fold change and are candidate RBPs, with three GRPs denoted. (B) Western blot using protein lysates from root hair and nonhair cell nuclei measuring GRP7 and GPR8 protein levels. There is no noticeable difference between cell types when normalized to the H3 loading control. See **Figure 3.10B** for the H3 loading control results. (C) RIP-qPCR was performed on whole root lysate using rabbit IgG (blue bars) or rabbit serum that recognizes native GRP7 and GRP8 (green bars) graphed as fold change relative to IgG. *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, Welch's t-test. Error bars indicate SEM.

As the GRP7/8-bound motif was enriched specifically in hair cell PPSs, we tested plants aberrantly expressing these proteins for root hair cell phenotypes. From this analysis, we found that root hair cell density in the *grp7-1* null mutant is significantly (*p* value < 3.3×10^{-7}; Wilcoxon test) increased compared to WT plants (**Figure 3.15A**). In accordance, plants overexpressing *GRP7* (*GRP7*ox) demonstrate significantly (p value < 2.8×10^{-8}; Wilcoxon test) decreased hair cell density compared to their respective WT plants (Col-2) (**Figure 3.15A**). As mentioned previously, GRP7 is known to bind to *GRP8* transcripts, thereby decreasing *GRP8* expression

levels (Schöning et al., 2008), resulting in the *grp7-1* and *GRP7*ox lines exhibiting significantly (*p* values < 0.05; Welch's t-test) increased or decreased *GRP8* RNA (**Figure 3.15B**) and protein (**Figure 3.15C**) levels as compared to WT plants, respectively. Thus, to differentiate the effects of each protein in hair cell differentiation, we required additional mutant plant lines. For instance, we identified a mutant line with an insertion in the *GRP8* promoter (CS803581/SAIL_75_G05; hereafter referred to as *GRP8*ox) that resulted in a significant (*p* value < 0.001; Welch's t-test) increase in the levels of *GRP8* mRNA and protein levels in these plants relative to WT. Importantly, this increase in *GRP8* levels does not cause a concomitant alteration in *GRP7* abundance in *GRP8*ox plants (**Figure 3.15B-C**). We examined root hair cell density in these plants, and revealed a significantly (*p* value < 0.015; Wilcoxon test) increased root hair density as compared to WT, strongly suggesting that this is a GRP8-dependent phenotype (**Figure 3.15A**). To determine the effects of altering GRP7 alone on root hair cell fate, we also measured the density of these cells in a plant line that contains a *GRP7* null mutation (*grp7-1*), as well as an artificial microRNA targeting *GRP8,* which returns the levels of this mRNA and protein close to those of WT (hereafter *grp7-1;8i*) (Streitner et al., 2012) (**Figure 3.15B-C**). We found that these plants exhibit a similar root hair density as WT (*p* value > 0.825; Wilcoxon test) (**Figure 3.15A**), indicating that this is indeed a GRP7 independent phenotype. Therefore, the *grp7-1* plants only exhibited increased root hair cell density as a result of increased *GRP8* levels, not due to the absence of *GRP7*.

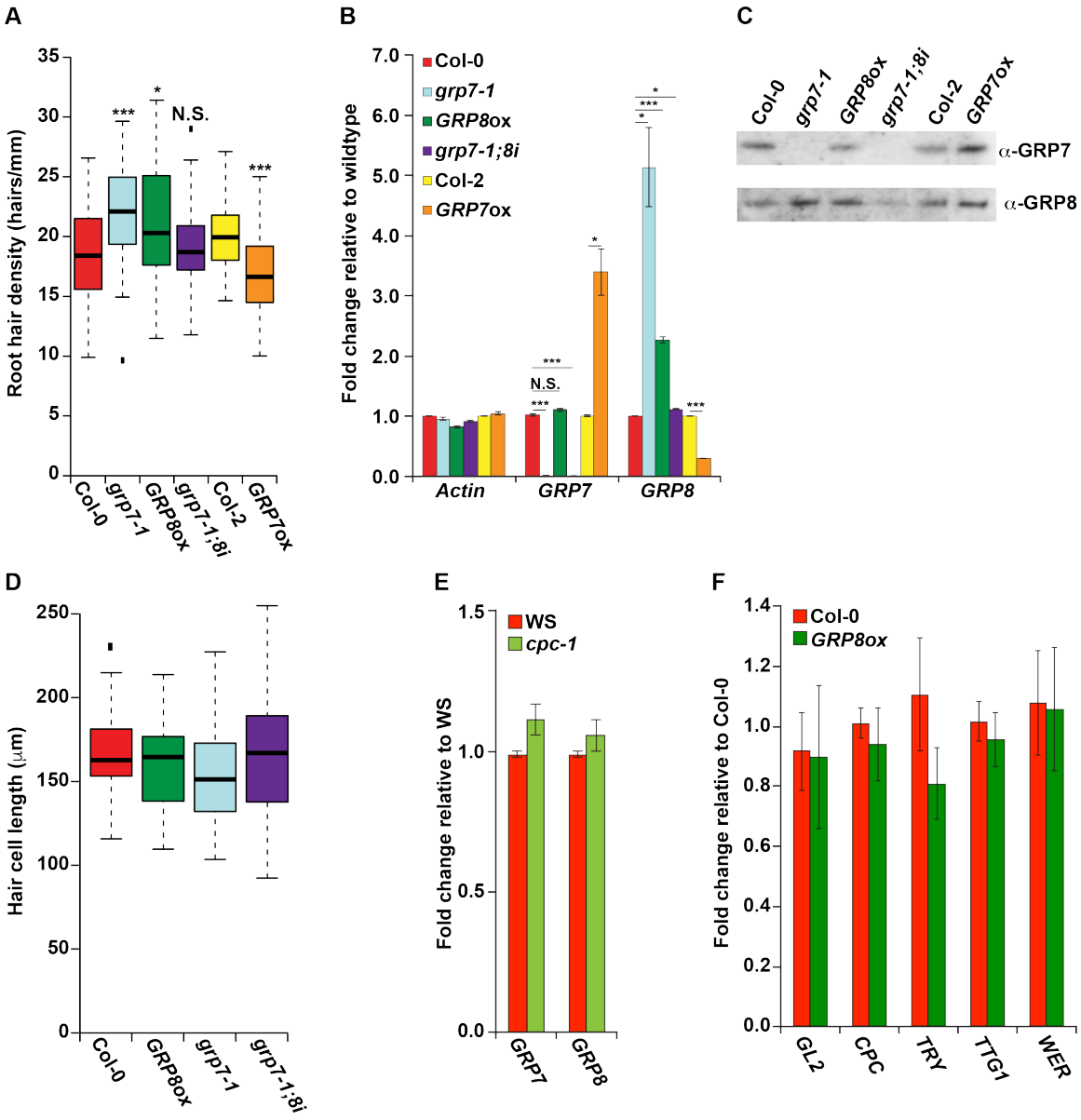


Figure 3.15: GRP8 regulates root hair cell fate in a GRP7- and CPC-independent manner
(A) Root hair cell density was measured in 8-day-old seedlings of WT or plants with decreased or increased *GRP7* (*grp7-1* or *GRP7ox*, respectively), increased *GRP8* (*GRP8ox*), or decreased *GRP7* with WT levels of *GRP8* (*grp7-1;8i*), n > 50. * and *** denote *p* value < 0.05 and 0.001, respectively, while N.S. denotes *p* value > 0.05, Wilcoxon test. (B) RT-qPCR of root tissue from lines with altered *GRP7* and/or *GRP8* levels, graphed as fold change relative to WT (Col-0 or Col-2). (C) Western blot measuring GRP7 and GPR8 levels in whole root lysate from twenty 8-day-old seedlings. (D) Measurement of the lengths of root hair cells in plants with altered *GRP7* or *GRP8* levels. The results show similar epidermal cell lengths in all tested genotypes. (E) RT-qPCR based measurements of *GRP7* and *GRP8* levels in the roots of WS (red) and *cpc-1* (light green) plants reveals no significant difference in their abundance. (F) RT-qPCR of transcripts involved in the CAPRICE/WEREWOLF transcription factor network using RNAs from the roots of Col-0 (red) and *GRP8ox* (green) plants. For (B), (E), and (F) *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, Welch's t-test. Error bars indicate SEM.

We next confirmed that this phenotype was due to ectopic hair cell production, rather than changes in the size of epidermal hair cells (**Figure 3.15D**). Like our analysis of SE, we probed GRP7 and GPR8 for a potential role in the CAPRICE/WEREWOLF pathway using qPCR analyses. Similar to the results for *SE*, we observed no significant (*p* value > 0.05; Welch's t-test) difference in *GRP7* or *GRP8* abundance in *cpc-1* roots compared to the wildtype control (**Figure 3.15E**). Additionally, we found no significant (*p* value > 0.05; Welch's t-test) difference in the mRNA levels of transcription factor that function in the CAPRICE/WEREWOLF pathway in the *GRP8*ox line compared to control (**Figure 3.15F**). Combined, these data reveal that GRP8 promotes root hair cell fate in a GRP7 independent manner, uncovering another novel post-transcriptional regulator of this important plant developmental process that is not affected by the CAPRICE/WEREWOLF pathway.

3.2.5 GRP8 promotes phosphate starvation stress response

One of the major factors regulating root hair cell fate is environmental signaling, such as nutrient deprivation. Therefore, a regulator of root hair cell fate may play a role in nutrient stress response. In fact, a recent microarray analysis of phosphate starved *Arabidopsis* roots revealed a mild increase in *GRP8* levels during the phosphate starvation response (Woo et al., 2012). Given this observation, in conjunction with our identification of GRP8 as a regulator of root hair cell fate (**Figure 3.15**), we next investigated the role of this RBP in the phosphate starvation stress response pathway. To begin, we performed RT-qPCR on the roots of WT plants grown on control and low phosphate media and validated that *GRP8* abundance is significantly (*p* value < 1.1×10^{-9}; Welch's t-test) upregulated upon phosphate starvation (**Figure 3.16A**), thereby verifying that this gene does respond to phosphate deprivation. We then examined the response of WT, *GRP8*ox, and *grp7-1;8i* plants to phosphate starvation. Using these plants, we first measured the levels of acid phosphatase activity from their roots under control and 3-day phosphate starvation conditions. This analysis revealed acid phosphatase levels to be significantly (*p* value < 0.05; Wilcoxon test) increased in the *GRP8*ox plants as compared to WT (**Figure 3.16B**) with no

significant (*p* value < 0.05; Wilcoxon test) difference between *grp7-1;8i* and WT plants (**Figure 3.16B**). These results indicate that there is a GRP8-dependent and GRP7-independent increase in acid phosphatase activity in *Arabidopsis* roots upon phosphate starvation.

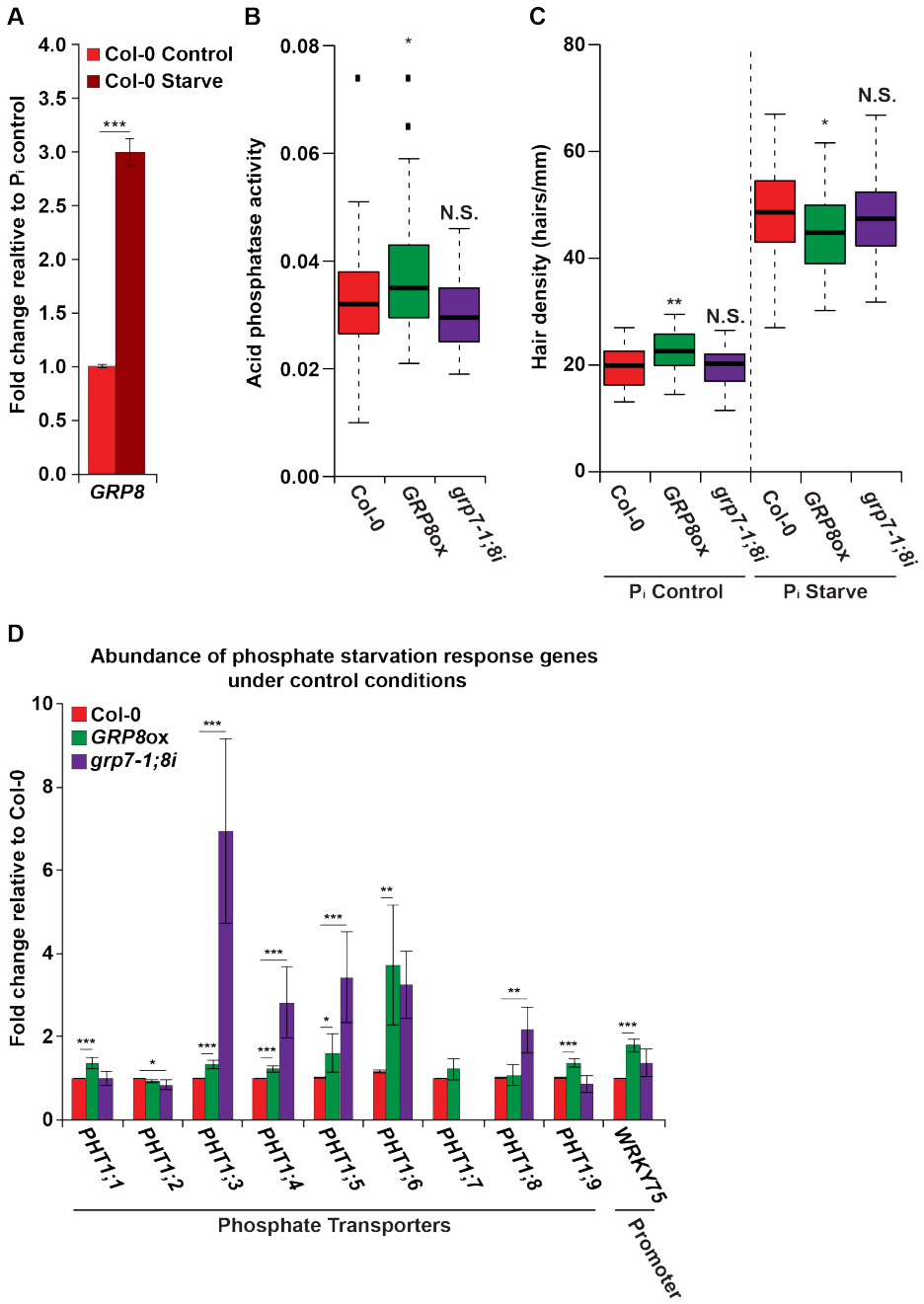


Figure 3.16: GRP8 functions in the phosphate starvation response pathway
(A) RT-qPCR measuring *GRP8* levels in Col-0 plants after three days of phosphate deprivation (dark red bar) or control treatment (light red bar). (B) Acid phosphatase activity in the roots of phosphate starved Col-0 and *GRP7/8* mutant 8-day-old seedlings, n > 40. (C) Root hair cell density (hairs/mm) in 8-day-old seedlings after three days of phosphate starvation. (D) Levels of

phosphate starvation response genes as measured by RT-qPCR in roots from Col-0 (blue), *GRP8ox* (green), and *grp7-1;8i* (purple) grown under control conditions. For (A) and (D), * and ** denote *p* value < 0.05 and 0.01, respectively, Welch's t-test. Error bars indicate SEM. For (B) and (C), * and ** denote *p* value < 0.05 and 0.01, respectively, Wilcoxon test.

Acid phosphatases are secreted from the root epidermis, therefore phosphatase activity corresponds to root surface area (Gilbert et al., 1999). To determine whether increased phosphatase activity is a consequence of increased root hair cell number, we measured hair cell density under both normal and phosphate deprivation conditions. From this analysis we observed that *GRP8*ox plants exhibited significantly (*p* value < 0.05; Wilcoxon test) decreased hair cell density under the starved conditions as compared to WT, while there was no change in *grp7-1;8i* plants (**Figure 3.16C**), indicating that there is an uncoupling of GRP8-dependent regulation of cell fate decision from phosphate starvation response. Furthermore, these findings demonstrate that the increase in acid phosphatase activity is especially sizeable in *GRP8*ox plants (**Figure 3.16B**), as there are fewer hair cells to secrete these enzymes during phosphate deprivation.

In order to better understand the roles of GRP8 and GRP7 in phosphate deprivation response, we measured the expression of numerous phosphate starvation response genes in the roots of *GRP8*ox and *grp7-1;8i* plants (Péret et al., 2011). To do this, we collected RNA from the roots of 8-day-old WT, *GRP8*ox*,* and *grp7-1;8i* seedlings under both control and phosphate starvation conditions and performed RT-qPCR on a number of *PHOSPHATE TRANSPORTER 1 (PHT1)* family genes (**Figure 3.16D**). We observed a significant (all *p* values < 0.05; Welch's t-test) increase in the levels of several *PHT1* family genes in the roots of *GRP8*ox plants under normal growth conditions (**Figure 3.16D**). PHT1 family genes are normally upregulated under phosphate starvation (Muchhal et al., 1996), providing a mechanism to maximize the uptake of phosphate when it is most scarce, allowing alleviation of the stress that the plant undergoes (Muchhal et al., 1996). Specifically, *PHT1;1* (*AT5G43350*) expression, which we found was increased in *GRP8*ox plants, has been linked to increased phosphate uptake and increased plant survival under phosphate starvation (Wang et al., 2014). In addition to heightened *PHT1* levels, we found significantly (all *p* values < 0.05; Welch's t-test) increased levels of the WRKY-domain containing transcription factor WRKY75 (AT5G13080) in *GRP8*ox as compared to WT roots

(**Figure 3.16D**). This is notable because WRKY75 is known to promote *PHT1;1* transcription during phosphate starvation, and may be involved in the transcription of other *PHT1* family genes (Wang et al., 2014). Interestingly, the *grp7-1;8i* plants exhibit upregulation of several PHT1 family genes (*PHT1;3, PHT1;4, PHT1;5, PHT1;8*) (**Figure 3.16D**), indicating that there is a GRP7-dependent inhibition of several of these genes. Therefore, these data indicate that there is a GRP8-dependent increase in the levels of most *PHT1* family transcripts, while GRP7 also affects several of these mRNAs.

We next aimed to determine if GRP8 directly binds to any of these phosphate deprivation response transcripts. As the α-GRP7/8 antibody binds to both GRP7 and GRP8 proteins, testing direct binding of GRP8 required performing RIP-qPCR in the roots of *grp7-1* plants grown under phosphate deprivation. Using this assay, we tested for GRP8 binding to *PHT1* family genes and *WRKY75*. Although there is no significant (all p values > 0.05; Welch's t-test) enrichment of *PHT1* family transcripts in GRP8 pulldown samples, we did observe a significant (all p values < 0.05; Welch's t-test) enrichment of *WRKY75* (**Figure 3.17A**; p value < 0.05; Welch's t-test) specifically in our α-GRP8 samples as compared to our IgG negative control. These data reveal that GRP8 binds to *WRKY75 in vivo*, leading to its altered transcript level. Thus, the GRP8-dependent regulation of *WRKY75* results in increased *PHT1* family phosphate transporter mRNA expression in the roots of 8-day-old seedlings.

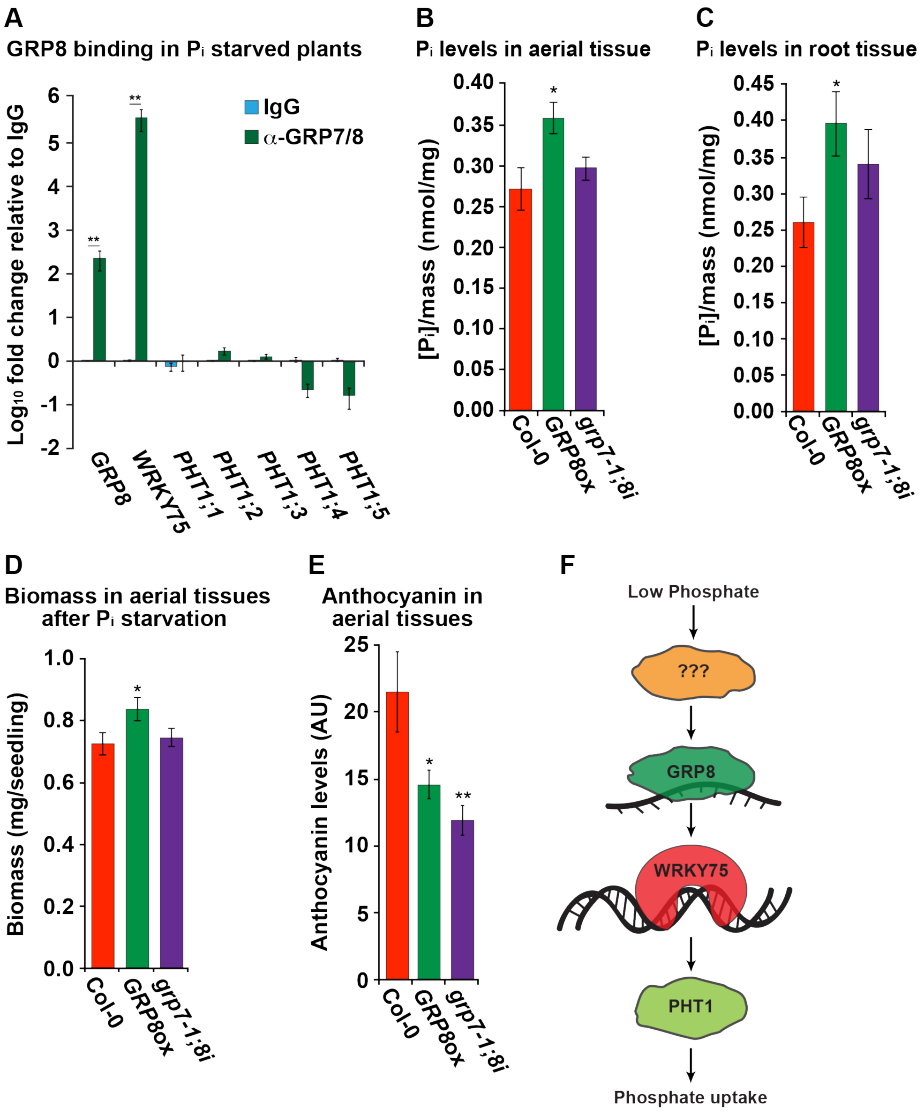


Figure 3.17: GRP8 alleviates phosphate deprivation stress
(A) RIP-qPCR of root tissue from *grp7-1* plants grown under phosphate starvation. RIP-qPCR was performed with a rabbit IgG (blue) or rabbit serum raised against GRP7 and GRP8 (green) graphed as fold change relative to α-IgG, n = 4 (B-C) Measurement of phosphate levels normalized to mass after 3-days of phosphate starvation in the shoots (B) or roots (C) of 8-day-old seedlings, n = 12. (D-E) Biomass (D) or anthocyanin levels (E) for 18-day-old seedlings after 2 weeks of phosphate deprivation, n = 12. For (A-E), *, **, and *** denote *p* value < 0.05, 0.01, and 0.001, respectively, Welch's t-test. Error bars indicate SEM. (F) A model of the role of GRP8 on the plant phosphate starvation response.

As GRP8 promotes phosphate transporter abundance, we next tested its role in alleviating both short-term and long-term phosphate starvation. We first measured phosphate levels in the aerial and root tissue of WT, *GRP8*ox, and *grp7-1;8i* seedlings after three days of phosphate starvation. This assay revealed significantly (*p* value < 0.05; Welch's t-test) increased

phosphate levels in both tissues in *GRP8ox* plants as compared to WT and *grp7-1;8i* seedlings (**Figures 3.17B-C**). These results indicated that both phosphate uptake and phosphate efflux to the shoots are upregulated in plants with higher *GRP8* levels. Additionally, we subjected plants to long-term (12-day) phosphate starvation and assayed both biomass and anthocyanin levels in the shoots of WT, *GRP8*ox, and *grp7-1;8i* seedlings, since phosphate starvation inhibits plant growth while promoting production of anthocyanin. We observed significantly (p value < 0.05; Welch's t-test) greater biomass in the shoots of *GRP8*ox as compared to WT and *grp7-1;8i* plants (**Figure 3.17D**). We also found significantly (p value < 0.05; Welch's t-test) decreased anthocyanin accumulation in the aerial tissue of both *GRP8*ox and *grp7-1;8i* as compared to the WT plants (**Figure 3.17E**). These data indicate that GRP8 is required for alleviating this plant stress by promoting increased phosphate uptake and biomass accumulation, while both GRP7 and GRP8 function in the reduction of the anthocyanin accumulation associated with phosphate starvation.

3.3 DISCUSSION

Here, we use PIP-seq to examine both the RNA-protein interaction and RNA secondary structure landscapes of nuclei from root hair and nonhair cells, which comprise the *Arabidopsis* root epidermis. Analyzing highly pure populations of hair or nonhair cell nuclei (**Figure 3.1**) revealed thousands of cell type-specific protein-bound sites as well as many shared sites, which are enriched in the coding sequence of the mRNA transcriptomes of both cell types (**Figure 3.6**).

This study compares global patterns of RNA secondary structure and RNA-protein interactions across the nuclear transcriptomes of two distinct cell types. This large-scale analysis identifies distinct profiles in specific regions of mRNA transcripts. For instance, mRNAs found in both cell types exhibit an increase in protein binding in the CDS, which corresponds to a relative decrease in secondary structure (**Figures 3.7A-B**). Interestingly, both RNA secondary structure and protein binding exhibit distinct patterns in the 3' UTR of root hair and nonhair cell nuclei (**Figures 3.8A-B**). RNA-protein interactions are known to be key regulators of numerous developmental events across various species. Therefore, our observation of cell type-specific

RNA-protein interactions is not very surprising. More interestingly, the main driver of RNA secondary structure has always been believed to be the primary sequence of a transcript. Thus, when analyzing the same RNA populations in two different cell types, one would expect to see virtually identical RNA folding patterns. Here, we describe distinct secondary structural profiles for mRNAs passing our expression threshold in both root hair and nonhair cell nuclei. These data reveal that there are further drivers of RNA folding in eukaryotic transcriptomes than simply the primary sequence, such as RNA-protein interactions, covalent modifications, and/or other factors. Furthermore, we use the cell type-specific patterns of RNA secondary structure and RNA protein interactions (**Figures 3.8A-B**) to identify the preferred interacting motif of GRP8 (**Figure 3.13**), which we then demonstrate is a novel regulator of root hair cell fate and plant phosphate stress response (**Figures 3.14-3.17**). Thus, our study exemplifies the use of a genome-wide screen to identify novel regulators of a biological process.

It is of note that we observe the greatest difference in both protein occupancy and RNA folding in the 3' UTRs of mRNAs expressed in both hair and nonhair cells (**Figures 3.8A-B**). These complementary profiles provide two potential models to explain this phenomenon. First, cell type-specific protein binding could regulate the folding of RNA transcripts, resulting in distinct folding patterns. Conversely, the distinct RNA folding patterns could in fact regulate protein binding. This latter model is supported by our findings that lncRNAs exhibit similar overall protein binding profiles while displaying distinct patterns of RNA secondary structure between root hair and nonhair cells (**Figures 3.8C-D**), suggesting that a different array of ssRNA- and dsRNA-binding RBPs are interacting with the distinctly structured lncRNAs found in these two cell types. Future studies will further investigate this type of cell type-specific RNA folding and RNA-protein interactions, determining the mechanisms involved in their feedback regulation.

This study also reveals an interesting pattern in nuclear RNA folding. Specifically, our analysis reveals that in both root hair and nonhair cell nuclei the CDS is less structured than both UTRs (**Figures 3.7A-B**), which is consistent with our nuclear PIP-seq performed in mixed nuclei from whole seedlings (Gosai et al., 2015). Although this pattern is consistent between all three

nuclear PIP-seq datasets, the opposite pattern has been observed in studies performed on whole cell (mostly cytoplasmic) RNA populations. These whole cell studies have been performed on unopened flower buds utilizing ds/ssRNA-seq (Li et al., 2012b), as well as on whole seedlings with structure-seq (Ding et al., 2014). Although these studies were performed using different techniques in a variety of *Arabidopsis* tissues, these data support the idea that the nuclear and cytoplasmic transcriptomes may in fact have distinct RNA secondary structure profiles. As with cell type-specific RNA folding, these distinct folding patterns could be due to different cohorts of RBPs in the nucleus and cytoplasm, and/or distinct post-transcriptional covalent modifications present in these cellular compartments. However, these consistent results across various studies and structure probing techniques warrant additional analyses to better understand this phenomenon.

In addition to describing global patterns, we used our PPS data to identify enriched protein-bound sequences and identify the RBPs that interact with a number of these sequences. More specifically, using RNA affinity chromatography we first identified SE as a candidate regulator of root hair cell development, while providing evidence of its preferred binding motif, a GGN repeat, in target RNAs (**Figures 3.10A-E**). Phenotypic analyses reveal that SE inhibits root hair cell fate in a miRNA biogenesis-dependent manner (**Figure 3.10F**), while also terminating root hair tip growth in differentiated cells (**Figure 3.10G**). We found that several SE-bound transcripts are necessary for proper root hair length (**Figures 3.12A-C**), and exhibit reduced abundance and stability in the absence of *SE* (**Figures 3.12E-F**). Combined, these data provide a working model in which SE functions with other microRNA biogenesis regulators to promote the nonhair cell fate, while also functioning independently to promote the stability of several mRNAs necessary for proper root hair termination (**Figure 3.12G**). Therefore, determining the mechanism by which SE positively regulates the abundance of these mRNAs that are required for proper root hair development is an area of future study.

Analyzing the first 100 nt of 3' UTRs, where the greatest difference in protein binding was observed, led to the identification of a hair cell-specific TG-rich motif (**Figure 3.13**). Through both

RNA affinity chromatography and phenotypic analyses we found that GRP8 binds this TG-rich motif, and promotes root hair cell fate in a GRP7- and CPC-independent manner (**Figures 3.14 and 3.15**). This finding is of particular interest since plants overexpressing GRP8 do not exhibit the deleterious aerial phenotypes described for *se-1* (Clarke et al., 1999; Serrano-Cartagena et al., 1999), making this a potential candidate for engineering more stress resistant crop plants. This idea is further supported by our observations that GRP8 is upregulated upon phosphate starvation, and promotes increased acid phosphatase activity (**Figure 3.16**). Additionally, we found that GRP8 alone has substantial effects in promoting phosphate uptake, efflux, and biomass accumulation while simultaneously alleviating anthocyanin production during phosphate starvation (**Figure 3.17**). In fact, our findings indicate the presence of a novel model of plant phosphate starvation response (**Figure 3.17F**). Specifically, we demonstrated that GRP8 is dramatically upregulated during phosphate starvation (**Figure 3.16A**), and promotes the abundance of mRNAs encoding phosphate transporters as well as *WRKY75*, which regulates transcription of several PHT1 transporters, while also binding directly to *WRKY75* (**Figures 3.17A-B)**. The increase in *PHT1* mRNA abundance likely explains the increased phosphate levels and biomass accumulation in *GRP8* overexpressing plants (**Figures 3.17C-D**), as well as decreased anthocyanin accumulation in the aerial tissues (**Figure 3.17E**). Thus, our working model suggests that phosphate deprivation initiates a signaling pathway that promotes *GRP8* abundance, which in turn binds to *WRKY75* mRNA thereby promoting its abundance. Increased *WRKY75* then promotes phosphate transporter expression, resulting in increased phosphate levels in the plant (**Figure 3.17F**). Although the specific signaling pathways and mechanisms by which GRP8 leads to increased *WRKY75* levels must be further elucidated, our findings point to the exciting possibility that *GRP8* overexpression is a viable target for engineering more stress resistant crop plants, which is a hypothesis that will be addressed with future research.

In total, our findings have revealed the power of PIP-seq in identifying biologically significant RBPs through a genome-wide screen. In the future, these newly described hair cell

regulatory proteins will be further studied to better understand the mechanisms by which they regulate this agriculturally important plant phenotype.

CHAPTER 4: COVALENT RNA MODIFICATIONS CORRESPOND TO TRANSCRIPT STABILITY

4.1 INTRODUCTION

Post-transcriptional regulation is an essential mechanism in controlling organismal development and response to external stimuli. Each RNA transcript undergoes a myriad of regulatory steps including splicing (Buratti and Baralle, 2004; Jin et al., 2011; Liu et al., 1995; Raker et al., 2009; Warf and Berglund, 2010), polyadenylation (Klasens et al., 1998; Oikawa et al., 2010), nuclear export (Grüter et al., 1998), subcellular localization (Bullock et al., 2010; Subramanian et al., 2011), translation (Kozak, 1988; Svitkin et al., 2001; Wen et al., 2008), and degradation (Goodarzi et al., 2012). Traditionally, RNA binding proteins (RBPs) have been the focus of study to understand post-transcriptional regulation, as these proteins perform the catalytic steps necessary for many of these events (Wahl et al., 2009). However, recent studies have shown that covalent nucleotide modifications play an essential role in many such regulatory events (Fustin et al., 2013; Wang et al., 2015; Xiao et al., 2016; Zhao et al., 2014; Zheng et al., 2013).

Each of the nucleotides comprising an RNA molecule is able to undergo covalent modification through the addition or restructuring of chemical groups, resulting in >100 chemically distinct nucleotides (Cantara et al., 2011; Limbach et al., 1994; Machnicka et al., 2013, 2013). Although the first modifications were discovered in the 1950s (Davis and Allen, 1957), both the localization and function of the majority of these modifications have remained a mystery. With the advent of next generation sequencing (Dominissini et al., 2012, 2016; Ryvkin et al., 2013; Schwartz et al., 2014b), we are now able to confidently identify the location of numerous modifications across the transcriptome, defining what has become known as the epitranscriptome (Meyer et al., 2012; Saletore et al., 2012).

One of the first techniques developed to study modification localization is the methy RNA immunoprecipitation and sequencing (meRIP-seq) technique. This technique was developed to identify multiple methylation modifications including N^6-methyladenosine (m^6A), N^1-methyladenosine (m^1A), methyl-5-cytosine (m^5C), and hydroxymethyl-5-cytosine (hm^5C) (Delatte et al., 2016; Dominissini et al., 2012, 2016; Hussain et al., 2013; Li et al., 2016; Meyer et al., 2012). In this assay, RNA is fragmented and an antibody targeting a specific modification of interest is used to immunoprecipitate fragments containing this modification. These RNA fragments are then used for high-throughput sequencing library processing, with a buildup of reads in the meRIP sample, compared to the negative control, indicating a site of modification. Each study utilizing meRIP-seq has revealed where along the transcriptome specific modifications are localized. In fact, this technique was used to correlate alternative splicing and transcript stability with the abundance of the m^6A modification (Xiao et al., 2016). The downside to this technique is that it is limited to studying modifications for which an antibody is available, and it requires generation of a separate sequencing library for each modification of interest.

Alternatively, the high-throughput annotation of modified ribonucleosides (HAMR) bioinformatics tool has been developed to overcome some of these drawbacks (Ryvkin et al., 2013). During sequencing library preparation, reverse transcriptase (RT) often stalls at the site of covalent modification, misincorporating a nucleotide at these positions, which appears as a buildup of mismatches from the reference genome in a next generation sequencing library. Using known modifications in a yeast tRNA training set (Ebhardt et al., 2009), the characteristic mismatch pattern was determined for numerous modifications. HAMR identifies high confidence mismatches across the transcriptomes, removing any mismatches that may be caused by sequencing errors or single nucleotide polymorphisms (SNPs). Additionally, HAMR has a validated machine-learning step that allows classification of eleven different modifications into seven distinct groups. Thus, HAMR can unambiguously identify pseudouridine (Ψ), dihydrouridine (D), 3-methylcytosine (m^3C), and 1-methylguanine (m^1G). Additionally, HAMR can identify, but not distinguish between isopentenyl adenine and N^6-isopentenyladenosine (i^6A | t^6A), 2-

methylguanosine and 2,2-dimethylguanosine (m^2G | $m^{22}G$), or between 1-methyladenosine, 1-methylinosine, and 2-methylthio-N^6-isopentenyladenosine (m^1A | m^1I | $ms^{2,6}A$) (Ryvkin et al., 2013). In addition to identifying known modification sites, HAMR is able to identify novel and verifiable sites of modifications (Ryvkin et al., 2013; Vandivier et al., 2015). Therefore, HAMR is a powerful bioinformatics tool able to be applied retroactively to any sequencing library in order to identify numerous modifications.

Studies utilizing either meRIP-seq or HAMR have correlated numerous post-transcriptional events with covalent modifications. Events such as alternative splicing (Xiao et al., 2016; Zhao et al., 2014; Zheng et al., 2013), translation efficiency (Wang et al., 2015), and RNA degradation (Vandivier et al., 2015) have all been linked to these modifications. Specifically, HAMR identifiable modifications have been found to be highly abundant on actively degrading mRNAs, indicating a link between covalent modification and transcript instability (Vandivier et al., 2015). Additionally, meRIP-seq experiments have shown distinct localizations of modifications; with m^1A being enriched near the start codon in human tissue culture cells (Dominissini et al., 2016; Li et al., 2016), while m^6A is enriched near the stop codon (Dominissini et al., 2012; Meyer et al., 2012). Furthermore, m^6A has been found to be essential for embryonic development in the model plant *Arabidopsis thaliana*, with mutants in the methyltransferase being embryonic lethal (Bodi et al., 2012). Therefore, previous studies have found examples of modification-specific localization (Dominissini et al., 2013, 2016; Schwartz et al., 2014b), modification-dependent post-transcriptional regulation (Dominissini et al., 2016; Xiao et al., 2016), and modification-dependent developmental phenotypes (Bodi et al., 2012; Zheng et al., 2013).

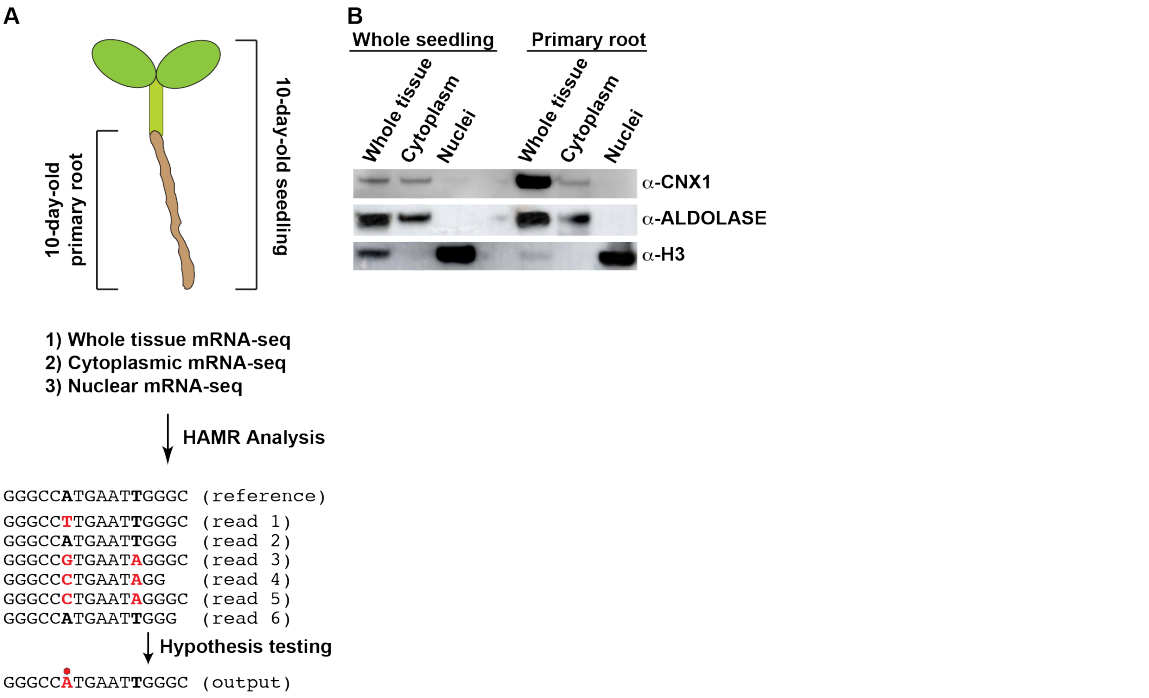
In this study, we performed a global analysis of covalent nucleotide modifications across the nuclear and cytoplasmic transcriptomes in both the *Arabidopsis* primary root and whole seedling tissue. These analyses revealed nuclear- and cytoplasmic-specific modifications, with each population having distinct localizations along the length of protein-coding mRNAs. Furthermore, we observed a strong correspondence between modification site and mRNA

stability. Combined, these data generate a working model whereby covalent modifications influence a variety of post-transcriptional processes across multiple *Arabidopsis* tissues.

4.2 RESULTS

4.2.1 mRNA-seq and HAMR analysis

In order to better understand the role of covalent modifications in post-transcriptional regulation, we produced next generation sequencing libraries to probe these modifications in whole tissue, nuclear, and cytoplasmic fractions of for both10-day-old *Arabidopsis* whole seedlings and primary roots (**Figure 4.1A**). To do this, we performed the isolation of nuclei in specific cell types (INTACT) nuclear purification protocol using ubiquitously tagged nuclei (Gosai et al., 2015; Wang and Deal, 2015). In this technique, a biotin receptor peptide is targeted to the nuclear envelope, allowing us to isolate highly purified nuclei lacking cytoplasmic markers (CNX1 and ALDOLASE) (**Figure 4.1B**). During INTACT the cytoplasmic fraction was also taken, and purity was verified by lack of the nuclear histone H3 (**Figure 4.1B**). Taking whole tissue, cytoplasmic, and nuclear fractions, we then performed polyA$^+$ selected RNA sequencing (mRNA-seq) on each of these samples. We performed 125 base pair paired end sequencing with a total of 17-73 million reads per library (**Figure 4.1C**). To ensure the reproducibility of these libraries we used DESeq2 to perform a clustering analysis using reads mapping to annotated mRNAs in all sequencing libraries generated from the whole seedling (**Figure 4.1D**) or primary root (**Figure 4.1E**) samples.



C

mRNA-seq library	Mapped reads
Seedling Whole Tissue Rep 1	28,415,887
Seedling Whole Tissue Rep 2	17,828,356
Seedling Cytoplasm Rep 1	73,558,722
Seedling Cytoplasm Rep 2	69,703,617
Seedling Cytoplasm Rep 3	68,452,564
Seedling Nuclei Rep 1	41,484,585
Seedling Nuclei Rep 2	68,096,467
Seedling Nuclei Rep 3	54,564,409
Root Whole Tissue Rep 1	30,767,888
Root Whole Tissue Rep 2	49,579,864
Root Whole Tissue Rep 3	39,404,913
Root Cytoplasm Rep 1	50,067,867
Root Cytoplasm Rep 2	36,415,111
Root Cytoplasm Rep 3	37,561,362
Root Nuclei Rep 1	22,227,536
Root Nuclei Rep 2	17,847,871
Root Nuclei Rep 3	52,113,621

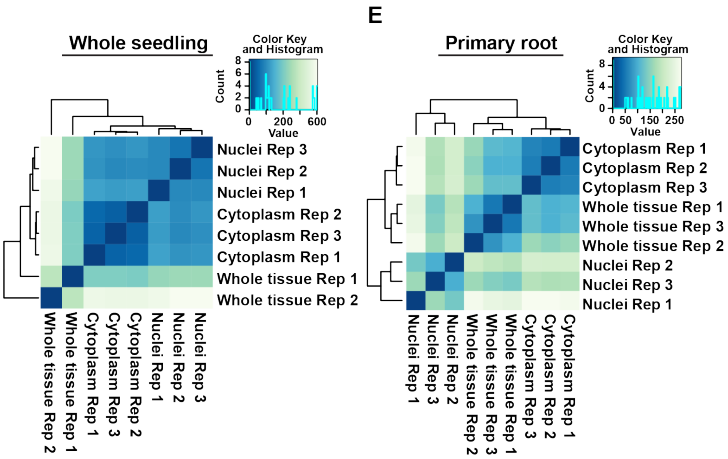


Figure 4.1: HAMR analysis of highly purified nuclear, cytoplasmic, and whole tissue transcriptomes.
(A) Diagram of experimental procedures. Taking 10-day-old whole seedlings or primary roots, whole tissue, cytoplasmic, and nuclear mRNA-seq was performed, followed by HAMR analysis. After mapping reads to the genome, HAMR can identify di-, tri-, and tetranucleotide mismatches (red text). Dinucleotide mismatches are discarded as possible SNPs, therefore only tri- and tetranucleotide mismatches undergo model tests to determine if they are modifications (indicated by a red hexagon). (B) Western blot of whole tissue, cytoplasmic, and nuclear fractions probing for the cytoplasmic markers CNX1 and ALDOLASE as well as the nuclear marker histone H3. (C) Table indicating the number of mapped reads per sequencing library. (D-E) DESeq2 clustering analysis of whole tissue, cytoplasmic, and nuclear libraries from whole seedling (D) or primary root (E) samples. Clustering analysis is based on the number of reads mapping to annotated genes.

Taking these mRNA-seq libraries we then performed HAMR analysis to identify covalent modifications (**Figure 4.1A**). We identified 14-2,400 covalent modifications in each library (**Figure 4.2**). Modifications had a modest to high reproducibility across biological replicates, ranging from 7.3% to 33.8% of modifications being present in at least two biological replicates (**Figures 4.2A-F**). We observed the highest reproducibility between samples with the fewest number of modifications (whole root and root cytoplasm), indicating this sequencing depth was able to identify a high proportion of modifications across the transcriptome (**Figures 4.2A and C)**. Additionally, we observed that samples with more modifications (root nuclei and all seedling samples) had a lower reproducibility, indicating that greater sequencing depth would reveal even more modifications (**Figures 4.2B and D-F**).

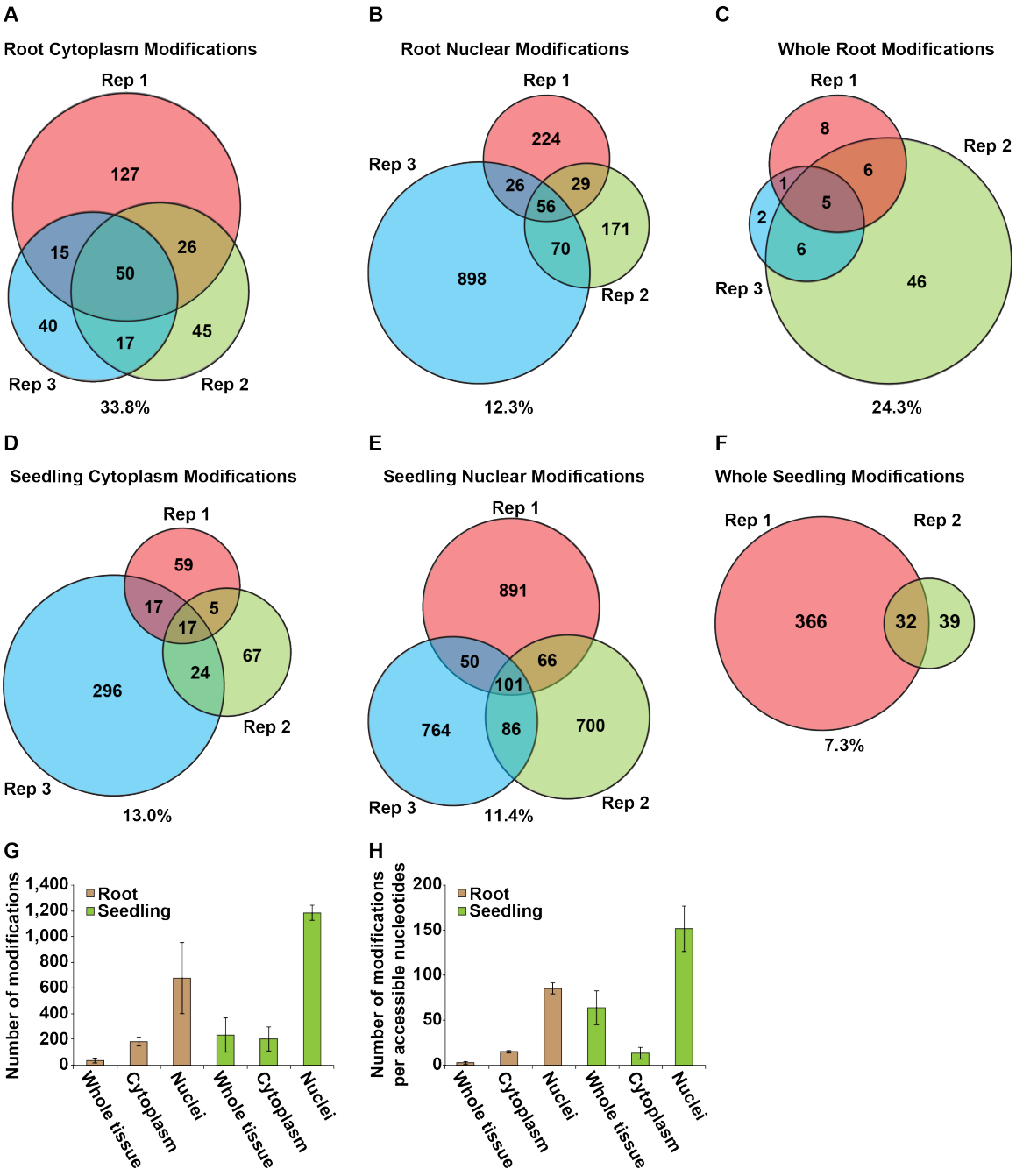


Figure 4.2: Covalent modifications are highly reproducible with higher abundance in the nucleus as compared to the cytoplasm of plant cells.
(A-F) Venn diagrams illustrating the reproducibility of individual modifications between biological replicates of the root cytoplasm (A), root nuclei (B), whole root (C), seedling cytoplasm (D), seedling nuclei (E), or whole seedling (F) libraries. (G) Number of modifications identified in each mRNA-seq library. (H) Number of modifications per accessible nucleotides in each mRNA-seq library. An accessible nucleotide is defined as any nucleotide with a read depth of at least 50, the minimum number of reads required by HAMR to call a modification. Error bars indicate standard error of the mean (SEM).

We next quantified the modifications identified in each sample. We examined both the absolute number of HAMR modifications, and the number of modifications per HAMR accessible bases (**Figures 4.2G-H**). During HAMR analysis we required a minimum depth of 50 reads to call a nucleotide as being modified, therefore our subsequent analyses only examine "accessible bases," those that have at least 50 reads of sequencing depth (Vandivier et al., 2015). Unsurprisingly, we observed that seedlings contain far more modifications than primary root tissue (**Figure 4.2G**). As covalent modifications can be cell type-specific (Batista et al., 2014; Chen et al., 2015), we expected to observe more modifications in seedlings, which have a higher number of distinct cell types than primary roots. Interestingly, we also found that the nuclear transcriptome is far more modified than the cytoplasmic transcriptome, whether examining either absolute number of modifications or modifications per HAMR accessible nucleotides (**Figures 4.2G-H**). To better understand this difference between nuclear and cytoplasmic modification abundance we next directly compared these epitranscriptomes.

4.2.2 The nucleus and cytoplasm have distinct epitranscriptomes

We first analyzed shared modifications at the single nucleotide level between nuclear and cytoplasmic libraries in both tissue types. We found that only 6.7% and 5.9% of nuclear modifications are present in the cytoplasmic transcriptomes of primary roots (**Figure 4.3A**) and seedlings (**Figure 4.3B**), respectively. Interestingly, we observed 16.3% of nuclear and 11.5% of cytoplasmic modifications are common to both the root (**Figure 4.3C**) and seedling (**Figure 4.3D**) transcriptomes. These data support previous studies that have described modifications as tissue or cell type-specific (Batista et al., 2014; Chen et al., 2015; Wan et al., 2015). As whole tissue samples are comprised of both nuclear and cytoplasmic transcriptomes, we expected to find a majority of whole tissue modifications to be present in our other samples. When examining modifications identified in whole root tissue we observed 68.4% of these modifications to be present in nuclear, cytoplasmic, or both libraries (**Figure 4.3E**). Surprisingly, we found only 8.6% of modifications from whole seedling tissue in either nuclear or cytoplasmic libraries (**Figure**

4.3F). This could be due to the far greater number of seedling modifications identified, indicating that greater sequencing depth would allow further discovery of modifications, thereby increasing this overlap.

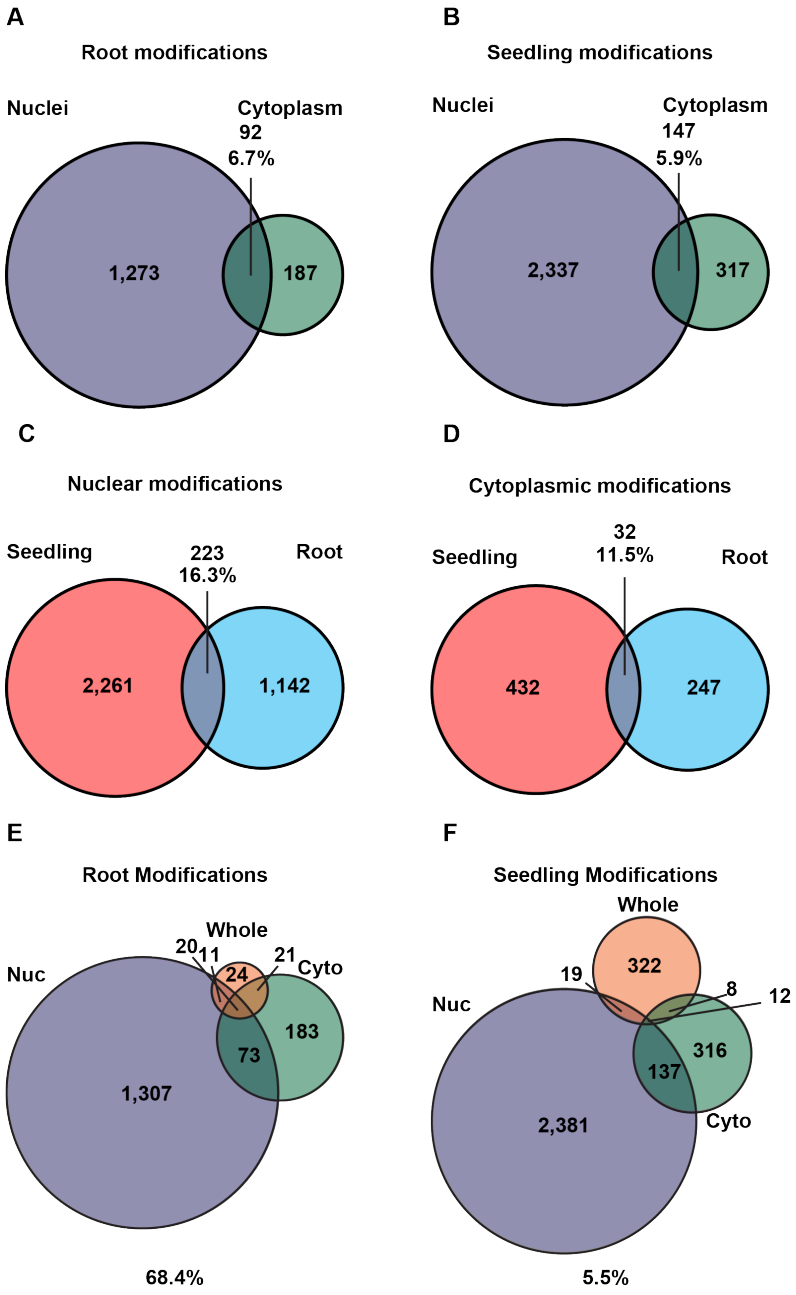


Figure 4.3: Covalent modifications are tissue and cellular compartment specific
(A-B) Venn diagrams indicating the percentage of nuclear (purple) modifications also present in the cytoplasm (green) in primary root (A) or whole seedling (B) samples. (C-D) Venn diagrams indicating the percentage of root (blue) modifications present in seedling (red) samples across both the nuclear (C) or cytoplasmic (D) transcriptome. (E-F) Venn diagrams indicating the

percentage of whole tissue modifications (orange) present in either the nuclear (purple), cytoplasmic (green) or both transcriptomes in primary root (E) or seedling (F) samples.

To further explore the possibility of distinct modification populations, we next analyzed the sites of these modifications across a transcript. We generated metagene profiles by dividing the 5' untranslated region (UTR), coding sequence (CDS), and 3' UTR into 100 equally sized bins, then graphing the percentage of genes with modifications occurring in each bin for primary root (**Figure 4.4A**) or seedling (**Figure 4.4B**) samples. This analysis revealed that modifications have distinct localizations along mRNA transcripts in the nuclear, cytoplasmic, and whole tissue transcriptomes. Interestingly, both tissue types displayed a similar distribution of modifications across mRNAs when comparing the same cellular compartments (e.g. nucleus to nucleus). For instance, we observed a vast majority of nuclear modifications (purple) in the 3' UTR, with a much smaller proportion of modifications occurring in the 5' UTR and CDS. Conversely, we observed a large proportion of cytoplasmic modifications (green) to be present in both UTRs. Additionally, modifications identified in whole tissue samples (orange) were most enriched in the 5' UTR with fewer observed in the CDS or 3' UTR. Together, these data support our hypothesis that the nuclear and cytoplasmic epitranscriptomes are distinct.

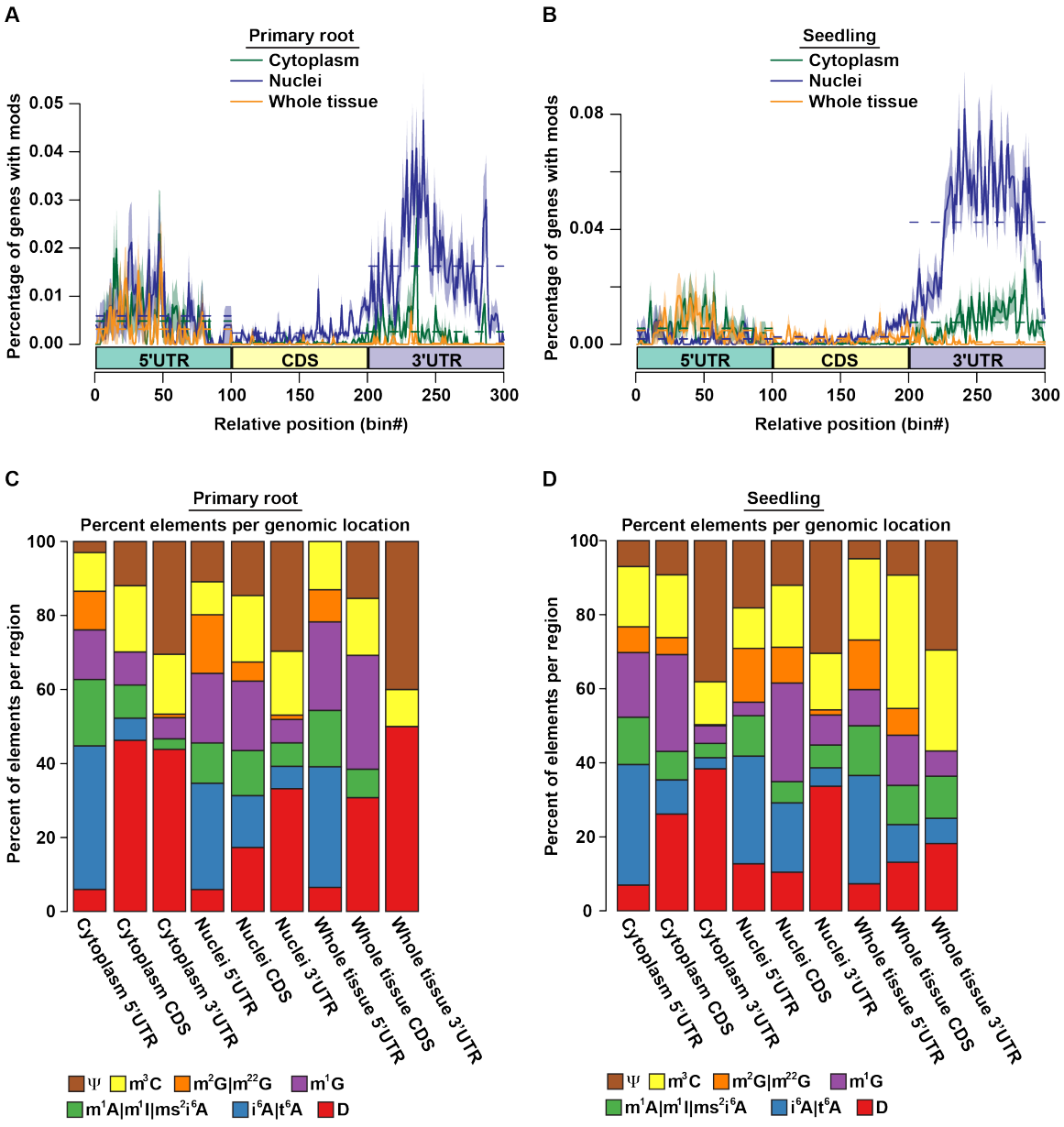


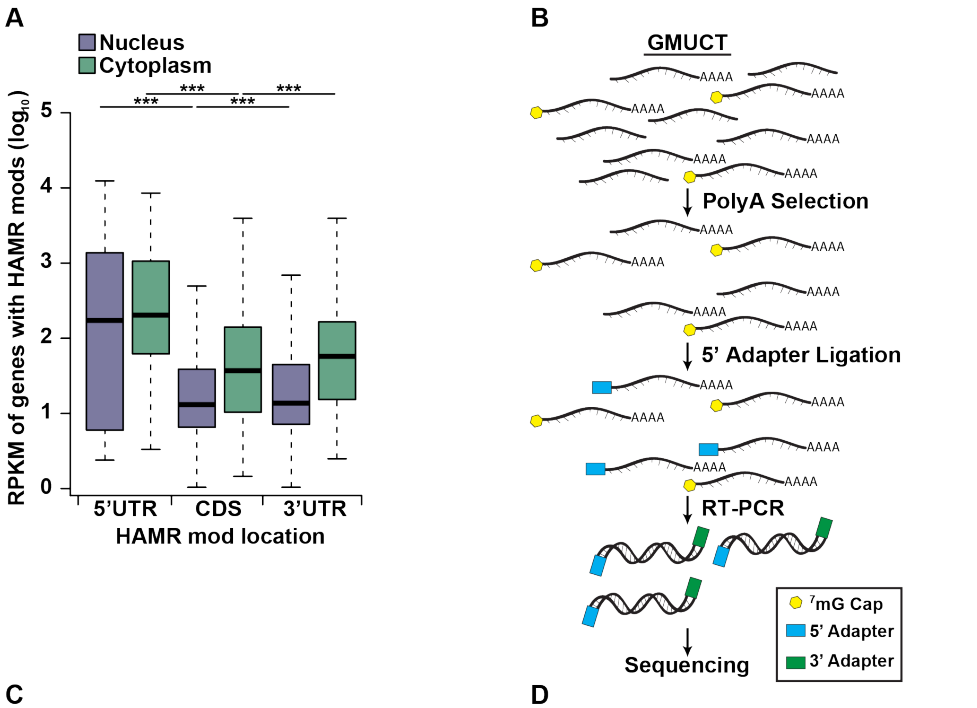
Figure 4.4: Nuclear and cytoplasmic epitranscriptomes have distinct modification localization and makeup

(A-B) Metagene profiles divide the 5' UTR, CDS, and 3' UTR of all detectable mRNAs into 100 equally sized bins. The percentage of genes with modifications that occur in each bin is graphed for the cytoplasmic (green), nuclear (purple), or whole tissue (orange) epitranscriptome in primary root (A) or whole seedling (B) tissue. Shading indicates the SEM. The dotted line indicates the mean percentage of modifications occurring across all 100 bins of the mRNA region. (C-D) A breakdown of modifications into the seven categories detectable by HAMR. Modifications are subdivided into those present in the 5' UTR, CDS, and 3' UTR of the cytoplasmic, nuclear, or whole tissue epitranscriptomes for primary root (C) or whole seedling (D) samples.

To further understand the different modification localization between intracellular compartments, we next classified the modifications identified. Using the ratio of mismatched reads, HAMR is able to confidently identify eleven different modifications, which it can assign into seven different groups (Ryvkin et al., 2013; Vandivier et al., 2015). We determined the identity of the modifications observed in the 5' UTR, CDS, or 3' UTR in all six samples, and graphed the percentage of total modifications that each class comprises in both primary root (**Figure 4.4C**) and whole seedling (**Figure 4.4D**) samples. We found that the 5' UTRs tend to have higher proportions of $i^6A|t^6A$ and $m^2G|m^{22}G$ with fewer Ψ and D than the CDS or 3' UTR. Interestingly, we also observed that only Ψ, D, and m^3C were identified in the 3' UTR of whole root tissue. Together these data support previous findings that modifications are mRNA location dependent, with specific modifications being enriched in various regions of the transcript.

4.2.3 Covalent modification site corresponds to transcript abundance and stability

As modification prevalence differs between transcript regions, we next aimed to determine the role of this localization on transcript abundance. To do so, we graphed the reads per million per thousand bases (RPKM) of genes with modifications in the 5' UTR, CDS, or 3' UTR (**Figure 4.5A**). We observed a significant (*p* value < 2.2×10^{-16}; Wilcoxon test) increase in the abundance of genes with modifications in the 5' UTR when compared to genes with modifications in the CDS or 3' UTR. Additionally, we observed that genes with modifications in the CDS are significantly (*p* value < 2.2×10^{-16}; Wilcoxon test) less abundant than those with modifications in either UTR. Interestingly, we also found that genes with nuclear modifications are consistently less abundant than those with cytoplasmic modifications (**Figure 4.5A**). This difference in abundance could be due to changes in RNA stability, which has previously been linked to transcript modification (Vandivier et al., 2015). Therefore, we next aimed to probe the stability of genes with modifications.



E

Region	*Mods*	*Cleaved mods*	*Percent*
mRNA	2,484	728	29.3%
5' UTR	44	0	0%
CDS	475	213	44.8%
3' UTR	1,590	465	29.2%

F

Region	*Mods*	*Cleaved mods*	*Percent*
mRNA	464	135	29.1%
5' UTR	72	0	0%
CDS	64	17	26.6%
3' UTR	294	87	29.6%

Figure 4.5: Modification site corresponds to mRNA abundance and stability
(A) Box plot indicating the $\log_{10}$ RPKM of genes exhibiting one or more modifications in the 5' UTR, CDS, or 3' UTR in the nuclear (purple) or cytoplasmic (green) transcriptome. (B) Schematic of GMUCT, in which RNAs undergo polyA$^+$ selection, followed by 5' adapter (blue box) ligation.

The 5' adapter can only ligate to mRNAs lacking a 5' cap (yellow hexagon), thereby allowing the subsequent amplification of mRNAs that have been uncapped or cleaved. The 3' adapter (green box) is then added during RT-PCR and samples are submitted for next generation sequencing. (C-D) Box plots indicating the proportion uncapped of genes with modified transcripts identified in the nuclear (C) or cytoplasmic (D) transcriptome. (E-F) Tables indicating the overlap of modification sites with GMUCT cleavage sites. Modifications are identified in nuclear (E) or cytoplasmic (F) mRNA-seq libraries while GMUCT was performed on whole seedlings. For (A), (B), and (D), *** indicates *p* value < 0.001 while N.S. indicates *p* value > 0.05, Wilcoxon test.

In order to assay actively degrading transcripts, we generated Global Mapping of Uncapped and Cleaved Transcripts (GMUCT) libraries using whole seedling tissue. In this technique, mRNA is isolated via polyA^{+} selection followed by 5' adapter ligation (**Figure 4.5B**). Using this method, only mRNA molecules lacking the 5' ^{7}mG cap, those transcripts that have been uncapped or cleaved, can be used as ligation substrates and undergo sequencing (Gregory et al., 2008; Willmann et al., 2014). It has been previously shown that mRNA abundance in GMUCT libraries is similar to mRNA-seq libraries (Gregory et al., 2008; Vandivier et al., 2015), therefore in order to assay mRNA stability the GMUCT samples must be normalized to total transcript abundance. The proportion uncapped is measured as the log ratio of the reads per million (RPM) of a gene as measured in GMUCT libraries, divided by the RPM of the same gene measured via mRNA-seq (Vandivier et al., 2015). This metric shows what proportion of the mRNA population for a gene is actively degrading.

In order to determine if modifications may affect transcript stability, we analyzed the proportion of uncapped transcripts for genes that contain or do not contain detectable mRNA modifications in the nuclear (**Figure 4.5C**) or cytoplasmic (**Figure 4.5D**) transcriptomes. This analysis revealed a significant (p value < 6.5 x 10^{-114}; Wilcoxon test) increase in proportion uncapped for genes containing mRNA modifications relative to those that lack modifications in the nucleus, indicating genes containing modifications are less stabile (**Figure 4.5C**). Interestingly, there is no significant (p value > 0.05; Wilcoxon test) difference in proportion uncapped between genes with or without cytoplasmic mRNA modifications (**Figure 4.5D**). To better understand this difference, we next examined the stability of genes with detectable mRNA modifications present in the 5' UTR, CDS, and 3' UTR. We found that genes that exhibit nuclear modifications in the 5' UTR are significantly (p value < 8.7 x 10^{-6}; Wilcoxon test) more stable than those exhibiting

modifications in other mRNA regions, with no significant (p value > 0.05; Wilcoxon test) difference between these and unmodified transcripts (**Figure 4.5C**). Similarly, we observed that genes with cytoplasmic modifications in the 5' UTR are significantly (p value < 1.6×10^{-6}; Wilcoxon test) more stable than genes without modifications (**Figure 4.5D**). Furthermore, we observed that genes with mRNA modifications in the CDS are significantly (p value < 1.4×10^{-11}; Wilcoxon test) less stable than other genes with mRNA modifications, while modifications in the 3' UTR indicating intermediate stability in both the nucleus and cytoplasm (**Figures 4.5C-D**). These data indicate that the localization of modification site may influence transcript stability.

To further examine the association between modifications and stability, we next examined the overlap between modification sites and cleavage events. In GMUCT, the 5' most nucleotide of a read corresponds the site of mRNA cleavage by endonucleases (Gregory et al., 2008). Therefore, we examined the overlap between whole seedling cleavage sites and modification sites found in the nucleus (**Figure 4.5E**) or cytoplasm (**Figure 4.5F**). We found a significant (p value < 2.2×10^{-16}; χ^2 test) enrichment for cleavage sites at the sites of both nuclear and cytoplasmic modifications compared to unmodified sites (**Figures 4.5E-F**), supporting earlier findings that modifications may signal for degradation (Vandivier et al., 2015). Furthermore, when examining modification sites found in the 5' UTR, CDS, and 3' UTR, we observed a striking correspondence between modification localization and cleavage sites. While 26-45% of modification sites overlap with cleavage sites, we observed no 5' UTR modification sites overlapping with cleavage sites (**Figures 4.5E-F**), significantly (p value < 0.002; χ^2 test) fewer than expected. These data further support our observations that the localization of the modification sites may influence transcript stability.

4.3 DISCUSSION

Here, we study the role of covalent modifications in post-transcriptional regulation of the nuclear and cytoplasmic transcriptomes. In this study, we performed next generation sequencing on polyA$^+$ selected RNA from whole tissue, nuclear, and cytoplasmic transcriptomes in both 10-

day-old seedlings and primary roots. These sequencing libraries were then analyzed using the HAMR bioinformatics pipeline in order to identify covalent modifications across the various transcriptomes (**Figure 4.1A**). From this analysis, we identified far more modifications in nuclear compared to cytoplasmic samples in both tissue types (**Figures 4.2G-H**), with very little overlap (**Figures 4.3A-B**). While numerous modifying enzymes are known to be nuclear (Brzezicha et al., 2006; Decatur and Schnare, 2008; Liu et al., 2014), the identity of many such enzymes is still unknown. The presence of cytoplasm-specific modifications supports the hypothesis that a subset of these unknown enzymes may in fact be localized to the cytoplasm. Surprisingly, we also observed thousands of nuclear-specific modifications. As many nuclear modified transcripts are believed to be exported to the cytoplasm, the absence of these modifications was unexpected, indicating the further processing of these mRNAs. These results lead to three possible explanations. 1) These modifications could be erased from mRNAs upon export from the nucleus. While many enzymes that remove modifications have been identified (Jia et al., 2011; Li et al., 2016; Zheng et al., 2013), the erasure of thousands of modifications is unlikely. 2) As modifications have been shown to influence nuclear export (Fustin et al., 2013; Zheng et al., 2013), nuclear-specific modifications could be signals for nuclear retention, with their removal allowing nuclear export. 3) These modifications could signal for degradation of these transcripts either before or immediately following nuclear export, explaining their absence in the cytoplasm. In this study, we tested the third hypothesis, however, some combination of two or more explanations is likely and will be need to be investigated further in the future.

To better understand the role of covalent modifications in post-transcriptional regulation we next examined the localization of these modifications. Previous studies of individual modifications have found distinct localizations for both m^1A and m^6A modifications (Dominissini et al., 2013, 2016). We generated metagene profiles to examine the localization of modification in the nuclear, cytoplasmic, and whole tissue transcriptomes (**Figures 4.4A-B**) and found distinct localizations between these nuclear and cytoplasmic modifications. This observation supports our hypothesis that the nuclear and cytoplasmic transcriptomes have distinct populations of

modifications. Specifically, while the nuclear modifications primarily localize to the 3' UTR, cytoplasmic modifications localize to both UTRs. As HAMR identifies eleven different modifications (Ryvkin et al., 2013), we next determined the makeup of modifications in each region of the mRNA, observing increased of $i^6A|t^6A$ and $m^2G|m^{22}G$ with decreased Ψ and D in the 5' UTR as compared to the CDS and 3' UTR in each of our six samples (**Figures 4.4C-D**). It is of particular interest that the 5' UTRs of nuclear mRNAs have this distinct makeup of modifications, providing a potential mechanism for marking the 5' UTR prior to nuclear export. These results further begged the question of what role these modifications play in post-transcriptional processing.

We next aimed to determine the role of nuclear and cytoplasmic modifications in mRNA stability. We first analyzed the correspondence between modification site and mRNA abundance. We observed that genes with mRNA modifications in the 5' UTR are more highly abundant than those with modifications in the CDS or 3' UTR (**Figure 4.5A**). We performed GMUCT on whole seedling tissue (**Figure 4.5B**) to quantify the proportion uncapped, or stability, of these modified mRNAs. We found genes with modifications mRNA in the CDS and 3' UTR to be significantly less stable than those with unmodified mRNAs, while genes with mRNA modifications in the 5' UTR exhibit no change in nuclear stability and increased cytoplasmic stability (**Figures 4.5C-D**). Furthermore, as GMUCT is able to find the exact site of mRNA cleavage (Gregory et al., 2008), we examined the overlap between RNA cleavage and modification sites. Interestingly, we observed a significant enrichment of cleavage sites at the site of modifications in the CDS and 3' UTR (**Figures 4.5E-F**). While previous studies have shown a strong correspondence between HAMR detectable modifications and transcript degradation (Vandivier et al., 2015), here we show that nucleotides that have the potential to be modified are far more likely to undergo cleavage than their unmodified counterparts. Although still correlative, this assay provides further support for the model by which covalent modifications are able to promote mRNA degradation, a model warranting further study.

Surprisingly, we observed zero mRNA cleavage sites that overlapped with sites of modification in the 5' UTR (**Figures 4.5E-F**). Combined with the fact that genes exhibiting these mRNA modifications have increased stability, these data suggest that covalent modifications in the 5' UTR uniquely correspond to more stable transcripts. This observation indicates that there are at least two distinct populations of covalent modifications, those that destabilize and those that stabilize mRNAs. It is of note that the modifications in the 5' UTR are comprised of a distinct modification makeup; therefore it is unclear whether modification identity, modification localization, or both factors are able to influence mRNA stability. Using covalent modifications, cells would be able to quickly respond to external stimuli by modifying the current population of RNA, promoting or inhibiting the stability of these mRNAs, thereby quickly altering the transcriptome before initiating a new transcriptional network. Further studies must examine changes in covalent modifications as initial responses to stimuli, prior to complementary changes in transcriptional output.

In total, this study has helped to illuminate the vast field of epitranscriptomics, and has raised many questions for future studies. We have identified multiple distinct populations of covalent modifications, with two distinct functions. Together, these findings help to illustrate a model in which modification identity and localization can lead to distinct processing fates for each mRNA transcript (**Figure 4.6**). Furthermore, as modifications can be made on the current transcriptome, this provides a mechanism for immediate response to signaling, far faster than altering transcriptional output. This variety of functions necessitates increased study of the epitranscriptome to more fully understand post-transcriptional regulation.

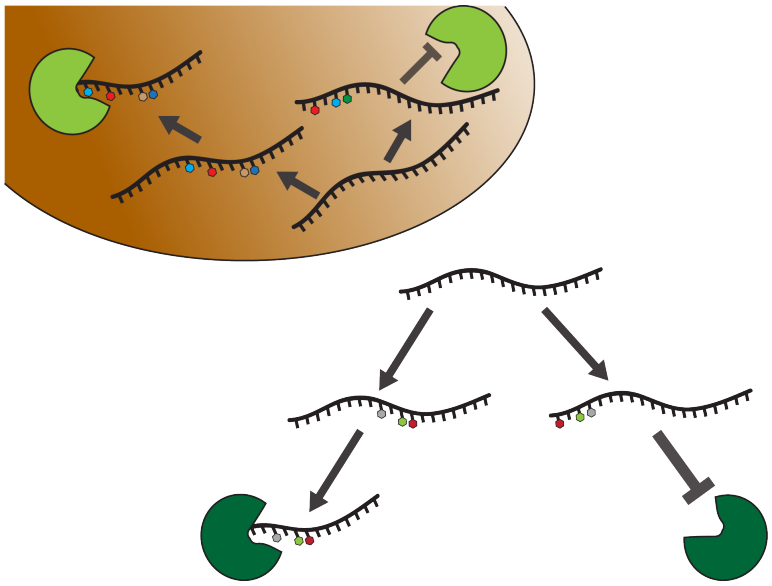


Figure 4.6: Working model illustrating the role of modifications in post-transcriptional regulation
This schematic illustrates how covalent modifications (colored hexagons) near the 5' end of a transcript inhibits degradation, while those in the CDS and 3' UTR may promote degradation machinery (green shapes).

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS

In this dissertation, I have examined the roles of both *cis* and *trans* acting post-transcriptional regulators in RNA processing. In Chapter 2, we performed the first simultaneous transcriptome-wide analysis of both RNA secondary structure and RNA-protein interactions in the nuclei of *Arabidopsis* seedlings. In Chapter 3, we utilized the protein interaction profile sequencing (PIP-seq) technique to identify RNA binding proteins (RBPs) that influence root hair cell development. Lastly, in Chapter 4, we examined the role of covalent nucleotide modifications in RNA stability. In this section, I will discuss the impact of these findings in a broader biological context, as well as potential directions for future studies to build upon this work.

5.1 NOVEL INSIGHTS INTO RNA-PROTEIN INTERACTIONS AND RNA SECONDARY STRUCTURE

5.1.1 Determining the regulators of nuclear-specific RNA secondary structure

In Chapter 2, we probed the RNA-protein interaction and RNA secondary structure landscape of the nuclear *Arabidopsis* seedling transcriptome. In this study, we observed increased RNA secondary structure in untranslated regions (UTRs) of mRNAs as compared to the coding sequence (CDS) (**Figures 2.7A-B**). This finding is particularly surprising as previous studies performed in whole seedlings and *Arabidopsis* flower bud whole tissue have shown the opposite trend (Ding et al., 2014; Li et al., 2012b). As this structural pattern was reproduced when performing PIP-seq on the nuclei of root hair and nonhair cells (**Figures 3.7A-B**), this may in fact be a nuclear-specific structural pattern.

These findings indicate the potential for nuclear and cytoplasmic transcriptomes to exhibit distinct structural profiles, indicating a massive restructuring of RNA upon nuclear export. To test this restructuring hypothesis, we performed PIP-seq to observe the native RNA structure in both the nuclear and cytoplasmic transcriptomes of 10-day-old *Arabidopsis* seedlings. From this analysis, we observed distinct nuclear and cytoplasmic secondary structure, with mRNAs exhibiting increased structure across the UTRs relative to the CDS in the nuclear transcriptome,

with the opposite pattern being present in the cytoplasm (**Figure 5.1A**). These results show a global restructuring of RNA upon nuclear export, begging the question of what regulates this restructuring. Altered RNA folding could be caused by differences in RNA-protein interactions between the two compartments, or through nuclear- and cytoplasm-specific covalent modification profiles, as observed in Chapter 4 (**Figures 4.4A-B**). To differentiate between these possibilities, we can perform the *in vitro* structure probing assay ds/ssRNA-seq (Li et al., 2012a, 2012b). In this technique, RNA is extracted from the cell and is then subjected to denaturing and reannealing, prior to treatment with single- or double-strand-specific RNases (**Figure 5.1B**). Therefore, by denaturing and reannealing the nuclear and cytoplasmic RNA in the absence of proteins, we can remove any protein-specific effects on RNA folding. Additionally, all covalent modifications will still be present in these samples. Thus, if covalent modifications are promoting changes in secondary structure then the nuclear and cytoplasmic *in vitro* structural profiles will mirror the native structures. Interestingly, we observed similar *in vitro* secondary structure profiles between nuclear and cytoplasmic mRNAs (**Figure 5.1C**). Therefore, these findings indicate that changes in RNA-protein interactions lead to a massive restructuring of mRNAs upon nuclear export.

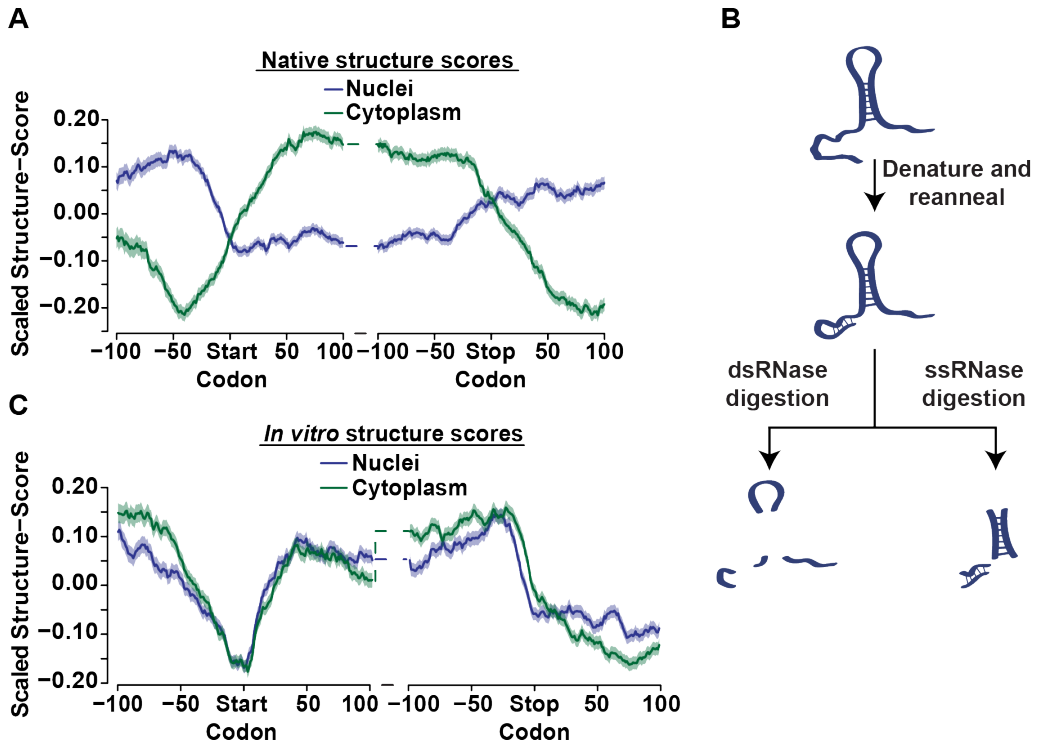


Figure 5.1: *Arabidopsis* mRNAs undergo a global protein-dependent restructuring upon nuclear export
(A) The native scaled structure score profiles for mRNAs from nuclear (purple) or cytoplasmic (green) fractions of 10-day-old *Arabidopsis* seedlings, at each nucleotide +/- 100 nt from the annotated start or stop codons. (B) Schematic of the *in vitro* structure probing assay ds/ssRNase-seq. Of note is the denaturing and reannealing step, in which the sample is heated and allowed to refold free of proteins, prior to treatment with ds- or ssRNases. (C) The *in vitro* scaled structure score profiles of nuclear (purple) or cytoplasmic (green) samples, at each nucleotide +/- 100 nt from the annotated start or stop codons. In (A) and (C), the solid lines indicate the mean scaled structure score, while the shading around the lines indicates the standard error of the mean (SEM).

This evidence of massive protein-dependent RNA restructuring raises numerous questions. Future studies must first investigate how exactly this restructuring is regulated, identifying specific RBPs that lead to altered RNA structure. Additionally, the role of this restructuring in mRNA processing must be further investigated. Studies should examine how a lack of restructuring would impact RNA stability, localization, and translation upon nuclear export. The field must address whether this restructuring is necessary for proper mRNA regulation, or simply an artifact of changing RNA-protein interactions upon nuclear export. These preliminary results lay the groundwork for years of study to better understand the role of RNA folding in post-

transcriptional processing.

5.1.2 Identifying post-transcriptional operons

During our nuclear PIP-seq analysis in Chapter 2, we identified highly bound RNA sequence motifs. We found that many of these motifs co-occur on numerous RNA transcripts with the same gene ontology terms (**Figure 2.12**), suggesting that these transcripts with similar function may be processed by the same cohort of RBPs. These results support the post-transcriptional operon hypothesis, which states that eukaryotic mRNAs with similar functions may contain identical *cis* regulatory sites thereby allowing processing by the same RBPs (Tenenbaum et al., 2011). This model allows numerous genes that function in similar pathways to be controlled as a single unit, analogous to a prokaryotic operon.

While evidence for post-transcriptional operons has been observed previously in yeast and mammalian samples (Keene and Tenenbaum, 2002; Silverman et al., 2014), this is the first description of such a regulatory mechanism in plants. Furthermore, these results show that PIP-seq is able to perform an unbiased screening of protein binding sites, which can be used to identify potential post-transcriptional operons. This analysis can be performed in agriculturally relevant systems to identify cohorts of protein-bound sequences in response to stimuli such as nutrient starvation, heat shock, or pathogen infection. In combination with RNA affinity chromatography, studies such as those performed in Chapters 2 and 3 can now identify novel regulatory networks through unbiased screens, providing new targets for genetic engineering of more stress resistant crop plants.

5.2 FURTHER INVESTIGATING THE ROLE OF RBPS IN ROOT HAIR CELL FATE

5.2.1 Utilizing PIP-seq to identify RBP regulators of development and stress response

In Chapter 3, we applied PIP-seq to an agriculturally relevant question, asking which RBPs function in proper root hair cell development. After performing PIP-seq on the nuclei of root hair and nonhair cells, we identified hair cell-specific, nonhair cell-specific, and common protein

protected sites (PPSs) (**Figure 3.5A**), and used these sequences to identify enriched protein-bound motifs (**Figures 3.9 and 3.13**). After identifying these motifs, we performed RNA affinity chromatography (**Figures 3.9, 3.10A, and 3.14A**) to identify proteins capable of binding these motifs. Together, this analysis allowed an unbiased screen to identify RBPs that had distinct cell type-specific functions.

While the work in this dissertation focused on examining two RBPs, SERRATE (SE) and GLYCINE-RICH PROTEIN 8 (GRP8), there are still 14 other RBPs identified as highly enriched in our RNA affinity chromatography (**Figure 3.9**). Future studies can begin examining the functions of these other RBPs in root hair cell development. Furthermore, there are another 42 protein-bound motifs that we have yet to use as baits in RNA affinity chromatography. Therefore, this single PIP-seq experiment has yielded a wealth of data that can be examined by our laboratory for years to come.

In addition to root hair cell development, this experimental paradigm can be applied to virtually any system to screen for RBPs involved in a biological process of interest. Now established, these protocols and bioinformatics pipelines can be applied to identify RBPs involved in the development of any plant cell type, plant response to stimuli, or many other stimuli. In these studies, PIP-seq can be used to compare two samples to identify RBPs involved in development, stress response, or many other stimuli. Furthermore, PIP-seq can be applied to other eukaryotic systems, allowing the study of human disease and development from a post-transcriptional perspective.

5.2.2 Determining the mechanism of SE-dependent hair cell regulation

In Chapter 3, we found that SE promotes nonhair cell fate in a microRNA biogenesis-dependent manner, and inhibits root hair length in a microRNA biogenesis independent manner (**Figures 3.10-3.12**). We observed that SE directly binds to *CAX4, MOR1,* and *PKL* mRNAs (**Figure 3.10E**), thereby increasing their abundance (**Figure 3.12E**) via increased stability (**Figure3.12F**). Together these data allowed us to generate a working model to describe the role

of SE in root hair cell development (**Figure 3.12G**). However, more work is necessary to fully understand the mechanisms by which SE functions in this developmental process.

As mutants in SE, ABH1/CBP80, and HYL1 all exhibit increased root hair cell density (**Figure 3.10F**) and function in conjunction to promote microRNA biogenesis (Dong et al., 2008; Yang et al., 2006), this is likely a microRNA biogenesis-dependent phenotype. Therefore, there are likely one or more microRNAs that promote root nonhair cell fate. Recent studies from the Benfey lab have been able to perform fluorescence activated cell sorting (FACS) on root hair and nonhair cells (Li et al., 2016). Although FACS does not generate samples as pure as the isolation of nuclei from specific cell types (INTACT) method (Deal and Henikoff, 2010), it is able to greatly enrich for whole cell samples. Therefore, FACS can be combined with small RNA sequencing (smRNA-seq) and Global Mapping of Uncapped and Cleaved Transcripts (GMUCT) analysis to identify hair and nonhair cell-specific microRNAs and their targets (Gregory et al., 2008). As we have found a clear link between microRNA biogenesis and root hair cell fate, studies such as these could determine the specific microRNAs and target mRNAs involved in this developmental process.

In addition to its role in promoting the nonhair cell fate, we have also observed a SE-specific termination of root hair length (**Figure 3.10G**). Recent studies have shown that SE has expanded function beyond its role in microRNA biogenesis, including as a regulator of alternative splicing (Laubinger et al., 2008; Raczynska et al., 2014). This work has shown that SE performs additional functions in promoting mRNA stability and root hair termination, making it a truly multifunctional RBP. The role of SE in root hair termination, and how its targets function to regulate termination, is of particular interest. Utilizing the *SE* hypomorphic mutant *se-1* (Clarke et al., 1999; Serrano-Cartagena et al., 1999), future studies can determine the SE-specific changes in mRNA stability or alternative splicing across plant root tissue. Additionally, *se-1* plants can be used in conjunction with FACS to enrich for hair and nonhair cells (Li et al., 2016), determining the role of *se-1* in alternative splicing and mRNA abundance in a cell type-specific manner.

Together, the identification of SE in both promoting nonhair cell fate and terminating root hair length opens this field to many new avenues of investigation.

5.2.3 Utilizing GRP8 in crop development

The final section of Chapter 3 focused on identifying the role of GRP8 in promoting root hair cell fate, and alleviating the effects of phosphate starvation. We first found that GRP8 functions independently of GRP7 to promote root hair cell fate (**Figures 3.15A-C**). When further investigating this function, we found that GRP8 binds to the mRNA of the phosphate starvation response regulator *WRKY75* (**Figure 3.17A**), leading to increased mRNA levels and subsequently increased phosphate transporter expression (**Figure 3.16D**). GRP8 overexpression serves to ameliorate the negative effects of phosphate starvation, leading to increased intracellular phosphate levels (**Figures 3.17B-C**), increased plant biomass (**Figure 3.17D**), and decreased anthocyanin accumulation (**Figure 3.17E**). Together these data provide a working model in which GRP8 leads to increased phosphate uptake, thereby reducing the stress of phosphate starvation (**Figure 3.17F**).

These findings are of particular interest as the GRP proteins are highly conserved across numerous plant species (Streitner et al., 2012). This work provides a novel approach to ameliorate the negative effects of phosphate starvation by overexpressing GRP8 in crop species. Work can be performed in other closely related plants within the genus *Brassicaceae* to determine whether GRP8 overexpression has comparable effects in agriculturally relevant dicotyledonous plants. Additional work is needed to further elucidate the mechanism of action by which GRP8 is connected to phosphate starvation. For instance, *GRP8* promoter-reporter gene fusion lines can be generated and subjected to random mutagenesis in order to identify genes that encode regulators necessary for GRP8 upregulation under phosphate starvation. Identifying these genes will further elucidate this novel stress response pathway, thus generating new targets for the potential genetic engineering of more phosphate starvation stress resistant crop species. Furthermore, identifying the role of other RBPs in phosphate starvation can reveal

additional signaling pathways, which may have analogous functions in crop species. Together, these studies will reveal target genes to allow increased agricultural yield and address increasing global demand.

5.3 DEFINING THE ROLE OF THE EPITRANSCRIPTOME IN RNA PROCESSING

5.3.1 Determining the mechanism of modification-dependent stability

In Chapter 4, we focused on examining the role of covalent nucleotide modifications in RNA processing. We observed that genes with covalent mRNA modifications also exhibit distinct mRNA stabilities (**Figure 4.5**). Specifically, we observed decreased mRNA abundance and stability for genes with mRNA modification in the CDS and 3' UTR (**Figures 4.5A-D**). Interestingly, we observed the opposite trend for genes containing mRNA modifications in the 5' UTR, exhibiting increased abundance and stability. Furthermore, while many sites of potential modification correspond to mRNA cleavage sites, there are no overlap between detectable 5' UTR modification and cleavage sites (**Figures 4.5E-F**). Interestingly, when examining the types of modifications present in the 5' UTR, CDS, and 3' UTR, we found a distinct modification makeup in the 5' UTR. Specifically, we observed increased levels of i^6A | t^6A and m^2G | $m^{22}G$, with decreased levels of Ψ and D (**Figures 4.4C-D**). Together, these data indicate that either 5' UTR modification localization or identity likely plays an important role in increasing mRNA stability.

Future studies are needed to determine the role of modifications in altered mRNA stability. While earlier studies have shown that actively degrading mRNAs are highly modification rich (Vandivier et al., 2015), this study correlates modification site with mRNA instability. Currently, these data are correlative. However, future studies can focus on identifying the mechanisms connecting covalent modifications and mRNA stability. Future studies can perform *in vitro* stability assays to determine whether covalent modifications directly regulate mRNA cleavage. Despite these limitations, however, these results reveal that the localization and/or identity of mRNA modifications correspond to the stability of a gene. Therefore, by performing a

single mRNA-seq experiment, researchers will be able to identify HAMR accessible modifications, and infer the relative stability of genes with mRNA modifications in the 5' UTR, CDS, and 3' UTR. These findings could greatly expand the information gathered from a single RNA-seq experiment, becoming a resource to the field at large.

5.4 CONCLUDING REMARKS

Post-transcriptional regulation allows the fine-tuning of the transcriptome to clearly define gene expression patterns in a cell type-specific manner. Furthermore, these regulatory events allow immediate responses to stimuli, altering the current transcriptome rather than waiting for the slower transcriptional machinery (Wahl et al., 2009). In Chapter 2, we have shown the correspondence between RNA secondary structure and RNA protein interactions, *cis* and *trans* acting regulators of RNA processing. In Chapter 3, we examined two distinct cell types to identify specific RBPs as novel regulators of root hair cell development. This study illuminated new functions of the RBP SE, and identified GRP8 as an agriculturally relevant regulator of root development and phosphate starvation response. Finally, in Chapter 4, we expanded our definition of *cis* acting features to include covalent modifications, examining the role of these features in mRNA stability. Together, this dissertation has illuminated a scarcely studied aspect of plant biology, identifying novel regulatory mechanisms and their interplay in post-transcriptional processing. These studies have developed the tools necessary for follow up work to better understand the role of *cis* and *trans* acting post-transcriptional regulators in numerous biological systems for many years to come.

APPENDIX A: MATERIALS AND METHODS

A.1 EXPERIMENTAL MODEL AND SUBJECT DETAILS

A.1.1 Plant materials

Seedlings were grown vertically on 0.5X MS plates with 1% sucrose and 0.8% Phytoblend, at 20°C, in a 16 h light/8 h dark cycle. The purified nuclei used in this thesis research were extracted from 10-day-old seedlings of *UBQ10*:NTF*/ACT2p*:BirA, or the primary root tissue of 10-day-old *ADF8*:NTF*/ACT2p*:BirA or *GL2*:NTF*/ACT2p*:BirA Columbia-0 (Col-0) ecotype of *Arabidopsis thaliana* using the INTACT methodology (Wang and Deal, 2015). All plants in this thesis research were ecotype Col-0 and grown under these conditions, unless otherwise noted. All plant genotypes can be found in **Table A.1**.

Plant Line	Source	Catalogue Number
UBQ:NTF/*ACT2p*:BirA	Deal and Wang, 2015	N/A
ADF8:NTF/*ACT2p*:BirA	Deal and Henikoff, 2010	N/A
GL2:NTF/*ACT2p*:BirA	Deal and Henikoff, 2010	N/A
Col-0	ABRC	CS70000
se-1	ABRC	CS3257
abh1-8	ABRC	SALK_117039
hyl1-5	ABRC	SALK_064863
WS	ABRC	CS28823
cpc-1 (WS ecotype)	ABRC	CS67760
cax4-1	Mei, H., *et. al*. 2009	N/A
mor1-1	ABRC	CS67061
pkl1-1	ABRC	CS3840
grp7-1	Streitner, *et al.,* 2012	N/A
Col-2	ABRC	CS69539
GRP7ox (Col-2 ecotype)	Streitner, *et al.,* 2012	N/A
GRP8ox	ABRC	CS803581
grp7-1;8i	Streitner, *et al.,* 2012	N/A

Table A.1: Plant lines used in this dissertation.
All plant lines used in this study are listed, including their source and catalogue number. Most lines are from the *Arabidopsis* Biological Resource Center (ABRC) and have a corresponding

catalogue number. Published plant lines that have not been deposited in ABRC have their first publication cited. All plant lines are Col-0 ecotype, unless otherwise noted.

A.2 METHOD DETAILS

A.2.1 Cross-linking and INTACT purification

Immediately before nuclei purification, plant tissue was crosslinked in nuclear purification buffer (20 mM MOPS, pH = 7, 40 mM NaCl, 90 mM KCl, 2 mM EDTA, 0.5 mM EGTA) plus 1% (vol/vol) formaldehyde under vacuum for 10 minutes, followed by a five minute quench with 125 mM Glycine under vacuum for an additional five minutes. Crosslinked seedlings then underwent INTACT purification as previously described (Wang and Deal, 2015).

Briefly, 3 g of root tissue or 1 g of seedling tissue was pulverized in liquid nitrogen, then resuspended in 10 mL of nuclear purification buffer. The resuspension was passed through a 0.45 μm nylon filter, and incubated on ice for 10 min. The samples were then centrifuged at 1,200 x *g* for 10 min, and the pelleted nuclei were resuspended in 1 mL of nuclear purification buffer. Following resuspension, 25 μL of streptavidin coated M-280 Dynabeads (Life Technologies; Carlsbad, CA, USA) were washed twice with nuclear purification buffer, and then combined with the samples in nuclear purification buffer. The beads were allowed to rotate end over end at 4°C for 30 min. Samples were then transferred to 15 mL conical tubes, and washed 4 times with 12 mL of nuclear purification buffer plus 0.1% Tween20. After the last wash the beads were resuspended in 1 mL of nuclear purification buffer and transferred to a 1.7 mL tube and washed twice more. The final samples were resuspended in 20 μL of nuclear purification buffer, snap frozen in liquid nitrogen, and stored at -80°C until processing.

A.2.2 Western blotting

Western blots using lysates from INTACT purified nuclei or 10-day-old roots were performed using α-ACT8 (1:5,000), α-PEPC (1:5,000; 200-4163S; Rockland; Boyertown, PA, USA), α-RUBISCO (1:5,000; ab62391; Abcam; Cambridge, MA, USA), α-BIP1 (1:200; sc-33757; Santa Cruz Biotechnology; Dallas, TX, USA), α-CNX1 (1:1,000; AS12 2365; Agrisera; Vännäs,

Sweden), α-H3 (1:1,000; ab1791; Abcam; Cambridge, MA, USA), or α-CP29A (1:5,000) α-EIF1A (1:1,000; AS10 934; Agrisera; Vännäs, Sweden), α-ALDOLASE (1:1,000; AS08 294; Agrisera; Vännäs, Sweden), α-SE (1:1,000; AS09 532; Agrisera; Vännäs, Sweden), α-ABH1/CBP80 (1:1,000; AS09 531; Agrisera; Vännäs, Sweden), α-AtGRP7 (1:1,000), α-AtGRP8 (1:1,000), or α-H3 (1:1,000; ab1791; Abcam; Cambridge, MA, USA) antibodies were performed as previously described (Kupsch et al., 2012).

Briefly, lysates were fractionated on a 4-12% SDS NuPage gel in MES at 100 V for 2 h. Transfer to PVDF was performed at 200 mA at 4°C for 2 h. The membrane was then briefly washed in water, and allowed to block at 4°C overnight in 5% milk in TBST. The membrane was blotted with primary antibody diluted in 5% milk at RT for 2 h, then underwent 3x 10 minute washes in TBST. The secondary antibody was diluted 1:5,000 in TBST and the membrane was blotted at RT for 2 h. Three 15 minute washes with TBST were performed, followed by one 5 minute wash in TBS. The membrane was then removed from liquid and ECL Prime Western Blotting Detection Reagent (GE Healthcare; Little Chalfont, UK) was applied to the membrane for one minute. Images were taken incrementally every 10 seconds until saturation of the images.

A.2.3 PIP-seq library preparation

~Two million INTACT purified nuclei were lysed in 850 μl RIP buffer (25 mM Tris-HCl, pH = 7.4; 150 mM KCl, 5 mM EDTA, pH = 7.5; 0.5% NP40; 10 μM DTT; 1 tablet protease inhibitors and 0.5μl/ml RNaseOUT (Life Technologies; Carlsbad, CA, USA)) by manual grinding. The resulting cell lysate was treated with RNase-free DNase (Qiagen; Valencia, CA, USA). The lysates were then split and treated with either 100 U/ml of a single-stranded RNase (ssRNase) (RNaseONE (Promega; Madison, WI, USA)) in 1X RNaseONE buffer for 1 hour at room temperature (RT), or 2.5 U/ml of a double-stranded RNase (dsRNase) (RNaseV1 (Ambion; Austin, TX, USA)) in 1X RNA structure buffer for 1 hour at 37°C as previously described (Silverman et al., 2014). Proteins were then denatured and digested by treatment with 1% SDS

and 0.1 mg/ml Proteinase K (Roche; Basel, Switzerland) for 15 minutes at RT. Proteinase digestion was followed by a 2 hour incubation at 65°C to reverse the RNA-protein cross-links.

To determine whether nuclease resistant regions in RNAs are due to protein binding or specific secondary structures, we also determined the digestion patterns of ds- and ssRNases immediately following protein digestion. To do this, we performed the identical treatments as described above except that the cross-linked nuclear lysates were treated with 1% SDS and 0.1 mg/ml Proteinase K (Roche; Basel, Switzerland) and ethanol precipitated prior to being treated with the two RNases. In this way, the SDS and Proteinase K solubilized and digested the proteins allowing us to deduce PPSs within all detectable RNAs in the cells of interest.

The digested RNA was then isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA, USA). To ensure that only high quality RNA samples were used for PIP-seq library preparation, the purified RNA was run on a Eukaryotic Total RNA Pico Series II chip (5067-1513; Agilent Technologies; Wilmington, DE, USA) using a BioAnalyzer 2100 system. Finally, the purified RNA was used as the substrate for strand-specific sequencing library preparation as previously described (Silverman et al., 2014). All of the RNase footprinting libraries (a total of 4 for each replicate: ss- and dsRNase treatments, footprint and structure only) were sequenced on an Illumina HiSeq2000 using the standard protocol for 50 base pair single read sequencing**.**

A.2.4 Total RNA sequencing library preparation to analyze the ratio of spliced to unspliced mRNAs

10-day-old seedlings of UBQ10:NTF/*ATC2p*:BirA underwent the INTACT purification as previously described (Deal and Henikoff, 2010). The resulting nuclei were lysed and the RNA was isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA, USA). Finally, the purified RNA was used as the substrate for strand-specific total RNA sequencing library preparation as previously described (Elliott et al., 2013), with the exception that no polyA$^+$ purification was performed, but was replaced by DSN treatments as

previously described (Silverman et al., 2014). The resulting libraries were sequenced on an Illumina HiSeq2000 using the standard protocol for 50 base pair single read sequencing.

6.2.5 UV Cross-linking analysis of motifs

Synthetic RNA oligonucleotides were radiolabeled in aT4 polynucleotide kinase (PNK) reaction (New England Biolabs; Cambridge, MA, USA) using 500 μCi of γ-^{32}P ATP following the manufacturer's recommendation, followed by phenol-chloroform extraction and precipitation. Each RNA probe was diluted to equal counts per minute (cpm), and was added to separate 10.2 μL binding reactions comprising 0.2 mM Tris pH = 7.5, 0.02 mM EDTA, 40 mM KCl, 1.3% polyvinyl alcohol, 25 ng/μl tRNA, 3 mM $MgCl_2$, 1 mM ATP, 50 mM creatine phosphate, and 2.8 μg/μl *Arabidopsis* leaf lysate in RIP buffer (25 mM Tris-HCl, pH = 7.4; 150 mM KCl, 5 mM EDTA, pH = 7.5; 0.5% NP40; 10 μM DTT; 1 tablet/10ml protease inhbitors) and incubated at 30°C for 20 minutes. The binding reaction was then subjected to UV cross-linking for 20 minutes using a 254nm UV lamp (Mineralight Lamp Model R-52G (UVP; Upland, CA, USA)). RNA-bound proteins were denatured in 1X SDS sample buffer and 1 mM β-mercaptoethanol and boiled for 5 minutes. Samples were separated on NuPAGE 3-8% Tris-Acetate gel (Life Technologies; Carlsbad, CA, USA) at 120V for 1 h. The gel was then fixed in a 10% methanol and 10% acetic acid solution for five minutes, and dried for 90 minutes. Phosphorimaging was used to visualize protein-bound and unbound RNA probes. This assay was replicated three times, and densitometry was used to quantify the bands that were present in both the motif and scramble probe lanes. The intensity of these bands was normalized to the intensity of the unbound probes from the corresponding lane, and the normalized intensity of the band in the scramble lane was set to one for comparison.

A.2.6 RNA affinity chromatography

We used motifs identified within PPS sequences as baits to isolate interacting proteins by affinity 'pulldown' studies. Specifically, RNA baits (covalently-linked to agarose beads) containing the identified motif of interest (IDT; Coralville, IA, USA) were incubated in a binding reaction (3.2

mM $MgCl_2$, 20 mM creatine phosphate, 1 mM ATP, 1.3% polyvinyl alcohol, 25 ng of yeast tRNA, 70 mM KCl, 10 mM Tris, pH 7.5, 0.1 mM EDTA) with ~60 μg of 10-day-old *Arabidopsis* lysate at RT for 30 minutes. Beads were washed four times with GFB-200 (20 mM TE, 200 mM KCl) plus 6 mM $MgCl_2$ and once with 20 mM Tris-HCl, pH = 7.4. The RNA-bound proteins were then directly trypsinized on the beads, as described below.

A.2.7 MS-ready sample preparation

Multiple independent samples for the selected motifs and their corresponding controls were used to average out experimental variability, optimize detection limits, and improve signal to noise ratio for robust specific identification. MS sample preparations and analyses were performed as described previously (Onder et al., 2008; Onder et al., 2006). Briefly, RNA-bound proteins were treated directly on the beads with 100 mM NH_4HCO_3 containing ~6 ng/μl of MS-grade trypsin (Promega; Madison, WI, USA) and incubated at 37ºC for 12-18 hrs. These samples were extracted first with 1%HCOOH/2%CH_3CN, and several times with 50% CH_3CN; combined peptide extracts were vacuum dried and desalted using a ZipTip procedure before resuspending in ~5-10 μL LC buffer A (0.1% HCOOH (v/v) in 5:95 CH_3CN:H_2O) for MS analysis.

A.2.8 RIP-RT-qPCR

RNA immunoprecipitation (RIP) was performed on whole leaf or whole root tissue from Col-0 or *grp7-1* as described previously. To begin, fresh tissue was submerged in PBS plus 1% (vol/vol) formaldehyde and vacuum infiltrated at room temperature (RT) for 10 minutes. One molar Glycine (Sigma-Aldrich; St. Louis, MO, USA) was added to a final concentration of 125 mM before an additional five minutes of vacuum infiltration. The tissue was then washed five times with distilled water, patted dry, and snap frozen in liquid nitrogen.

On the day of the RIP, the tissue was ground into a fine powder in liquid nitrogen using a mortar and pestle, and resuspended in RIP buffer (150 mM NaCl, 20 mM Tris, pH = 8.6, 1 mM EDTA, 5 mM $MgCl_2$, 0.5% NP40, 1 tablet/10 ml protease inhibitor (Roche; Basel, Switzerland),

0.5 μl/ml RNaseOUT RNaseOUT (Life Technologies; Carlsbad, CA, USA) at ~1 g/mL. This lysate was then subjected to 30 min of sonication (30 s on and 2 min off) and centrifuged twice for 15 min at max speed to remove any pelleted debris.

While the tissue is being prepared, 50 μL of Protein A beads (Life Technologies; Carlsbad, CA, USA) were washed twice with PBS then resuspended in 400 μL. Antibodies were then added to the beads at 5-10 μg per reaction, and allowed to rotated at 4°C for >2 hours. The antibodies used were α-CP29A, α-His, α-SE, α-ABH1/CBP80, rabbit serum raised against native recombinant *Sinapis alba* GRP10, which recognizes *Arabidopsis* GRP7 and GRP8 or normal rabbit IgG (3125, Cell Signaling Technology; Danvers, MA, USA). The beads were then washed twice with RIP buffer, and resuspended in plant lysate, followed by a 90 min rotation at 4°C. The RIP was then washed six times with RIP buffer, and resuspended in QIAzol. Immunoprecipitated RNA was then isolated using the miRNeasy mini kit (Qiagen; Valencia, CA, USA), and an RT was performed on 100-200 ng of RNA using Superscript II (18064014, Life Technologies; Carlsbad, CA, USA) with random hexamer priming, following the manufacturers protocol. For RIP-RT-qPCR performed on root tissue, the cDNA was subjected to 15 cycles of preamplification using the SsoAdvanced PreAmp Supermix (172-5160, BioRad; Hercules, CA, USA) kit, following the manufacturer's protocol. The template DNA was then used to perform qPCR using the 2x SYBR Green Master Mix (B21202, Bimake, Houston, TX, USA) and following the manufacturer's protocol.

A.2.9 Measurement of root hair density and root hair length

Seeds were sterilized in a 30% Clorox solution for 15 min followed by five washes with autoclaved water. After the last wash seeds were resuspended in 0.15% sterile agarose and stratified at 4°C for at least 48 hours. Seedlings were grown on 0.5X MS plates with 1% sucrose and 0.8% Phytoblend, grown vertically at 20°C, in a 16 h light/8 h dark cycle. Measurements of basal root hair density and length were performed on 8-day-old seedlings by imagine with a

dissecting microscope and measuring root hair length using JBrowse. Root hair density was calculated by measuring a length of primary root and counting all visible hairs along that length.

For phosphate starved plants, all seeds were planted on the described 0.5X MS plates and incubated for 5 days. On the fifth day the seedlings on each plate were transplanted to two new plates, one identical 0.5X MS plate and one 0.5X MS plate without phosphate. The control and starved plates were then replaced in the incubator for another three days. The root hair cell density and root hair length were then measured as described above.

For the temperature sensitive *mor1-1* plants, the plants were grown at 20°C for four days, then transferred to 31°C for another two days before imaging and phenotyping.

A.2.10 Measurement of RNA stability

To measure the stability of mRNA transcripts we took 8-day-old plants grown on 0.5X MS plates and excised the roots below the hypocotyl. Taking 30 roots per biological replicate, we then submerged these into 6 mL of liquid 0.5X MS media supplemented with 1% sucrose and 10 μM Actinomycin D. To account for any immediate changes to the transcriptome, we allowed these to incubate for 4 hours before taking our baseline 0 hour time point. We then collected roots after 8 hours of treatment, dabbing them dry of media, and snap freezing them in liquid nitrogen. We then proceeded to perform RNA extraction and RT-qPCR as previously described.

A.2.11 Measurement of acid phosphatase activity

To measure acid phosphatase activity, plants that had been phosphate starved were taken and the primary root was excised and placed in 300 μL of assay buffer (3.4 mM 4-naphthyl phosphate, 2.5 mM FastRed TR) and incubated at RT for 15 min. Then 150 μL of assay buffer was taken and absorbance at 405 nm was measured.

A.2.12 Measurement of phosphate concentration

Seedlings were germinated on 0.5X MS plates, and 5-day-old seedlings were transplanted to control or phosphate starved plates for three days. After phosphate starvation, the hypocotyl was cut to separate the seedlings into roots and shoots, and the tissue from five seedlings was pooled and weighed. This tissue was immediately placed into 1 mL of 1% glacial acetic acid and frozen in liquid nitrogen. The tissue underwent 8 rounds of freezing and thawing in liquid nitrogen and an RT water bath. After the eighth round of thawing, the samples were centrifuged at 20,000 x *g* for 5 minutes, and 100 μL of supernatant was taken and placed into 200 μL of water and 700 μL phosphate assay buffer (A: 2.85% H_2SO_4, 0.85% NH_4MoO_4, B: 10% ascorbic acid, A:B = 6:1). The samples were then incubated at 37ºC for 60 minutes, and absorbance was measured at 810 nm (Zhang et al., 2014). A standard curve was generated and the concentration of soluble phosphate per milligram of tissue was reported.

A.2.13 Measurement of anthocyanin

Seedlings were germinated on 0.5X MS plates, and 3-day-old seedlings were transplanted to control or phosphate starved plates for 14 days. After phosphate starvation, the hypocotyl was cut to separate the seedlings into roots and shoots, and the aerial tissue from five seedlings was pooled and weighed. The tissue was then submerged in a 18:1:81 solution of propanol:HCl:water, before incubation at 100ºC for 3 min. Samples were then centrifuged at >20,000 x *g* for 15 min. The supernatant was taken and absorbance was measured at 535 nm and 650 nm. The absorbance due to anthocyanin was calculated as: $A_{anthocyanin} = A_{535}-A_{650}$

A.2.15 mRNA-seq library preparation

Nuclear, cytoplasmic, or whole tissue RNA was extracted from whole seedling or primary root tissue, and resuspended at 20 ng/μL in water. Taking 50 μL of RNA (1 μg) we generated next generation sequencing libraries using the Illumina TruSeq stranded mRNA library

preparation kit with polyA$^+$ selection, following the manufacturer's protocol. Completed libraries underwent 125 bp paired end sequencing.

A.2.16 GMUCT library preparation

GMUCT libraries were constructed using RNA from 2 biological replicates of 10-day-old whole seedling using the GMUCT 2.0 protocol (Willmann et al., 2014). In brief, RNA was first subjected to polyA$^+$ selection followed by immediate ligation of a 5' RNA adapter. An additional polyA$^+$ selection step was performed to purify the adapter ligated RNAs. These samples were used as the substrates in reverse transcription reactions using a reverse primer that was composed of the 3' adapter sequence on the 5' end and a random hexamer on its 3' end. This allowed for the addition of the 3' sequencing adapter during reverse transcription. Finally, the GMUCT libraries were amplified and indices were added using a limited PCR amplification reaction.

A.3 QUANTIFICATION AND STATISTICAL ANALYSIS

A.3.1 Experiment specific information

The measurement precision, number of biological replicates (n), statistical tests performed, and significance for each experiment can be found in the figure legend of the experiment, as well as the RESULTS section of each chapter.

A.3.2 Read processing and alignment

All sequencing reads were first trimmed to remove 3' sequencing adapters using cutadapt (version 1.2.1 with parameters -e 0.06 -O 6 -m 14). The resulting trimmed sequences were collapsed to unique reads and aligned to the TAIR10 *Arabidopsis* genome sequence using TopHat (version 2.0.10 with parameters --library-type fr-secondstrand --read-mismatches 2 --read-edit-dist 2 --max-multihits 10 --b2-very-sensitive --transcriptome-max-hits 10 --no-coverage-search --no-novel-juncs). For 125 bp paired end analysis of mRNA-seq libraries 8 mismatches

were permitted. PCR duplicates were collapsed to single reads for all subsequent PIP-seq analyses.

A.3.3 Estimating unspliced transcripts

To estimate unspliced nuclear mRNAs, all reads from the total RNA-sequencing data that mapped to all detectable first TAIR10 annotated constitutively spliced introns were collected, removing reads that were entirely within the intron. We quantified the number of reads that had mapped through the exon/intron boundary (unspliced) compared to those that contained the exon/exon boundary (spliced). We then determined the fraction of junction mapping reads that were unspliced for each gene.

A.3.4 Identification of PPSs

PPSs were identified using a modified version of the CSAR software package (Muiño et al., 2011). Specifically, read coverage values were calculated for each base position in the genome and a Poisson test was used to compute an enrichment score for footprint versus structure only libraries. PPSs were then called with a false discovery rate (FDR) of 5% as previously described (Gosai et al., 2015; Silverman et al., 2014).

A.3.5 Functional analysis of PPSs

PPS annotation was done ‘greedily’ using the TAIR10 genome annotations, such that all functional annotations that overlapped with a given PPS were counted equally. Conservation was assessed by comparing both PhastCons scores and the number of SNPs, within PPSs relative to equally sized flanking regions. PhastCons scores for PPSs compared to same sized flanking regions were calculated as previously described (Li et al., 2012; Silverman et al., 2014).

To perform the SNP occurrence analysis we first identified SNPs located in transcriptionally active region (TARs), defined as intervals at least 15 nt long with greater than 20 reads of coverage, while allowing for a gap of 10 nt with less coverage, as calculated using an

aggregate list of alignments from both replicates of the PIP-seq libraries. Ten permutations of random shuffling of TARs were then performed to generate the control set with similar numbers and fragment sizes to our list of PPSs. We then quantified the number of non-redundant, substitution SNP sites cataloged by the 1001 Genomes Project (Cao et al., 2011) within the total list of PPSs and the 10 shuffled intervals, which were statistically compared to one another using a χ^2-test.

A.3.6 lincRNA conservation analysis

Brassica rapa lincRNAs were identified from a list of 3,450 intergenic transcripts generated previously (Tong et al., 2013), then further filtered by removal of transcripts with an open reading frame >100 codons. A total of 1908 *B. rapa* lincRNAs were then used as the dataset in BLAST analyses with *Arabidopsis* lincRNAs using an E-value of 10^{-10}.

A.3.7 Calculating the structure score statistic

For every base of detectable transcripts, we calculated the dsRNA-seq and ssRNA-seq coverages from the structure only samples, then calculated the structure score as described previously (Gosai et al., 2015; Li et al., 2012a, 2012b). Briefly, when given the dsRNA-seq and ssRNA-seq coverages (n_{ds}, n_{ss}) of a given base *i*, the structure score is determined as:

$$S_i = \mathrm{glog}\left(ds_i\right) - \mathrm{glog}\left(ss_i\right) = \log_2\left(ds_i + \sqrt{1+ds_i^2}\right) - \log_2\left(ss_i + \sqrt{1+ss_i^2}\right)$$

$$ds_i = n_{ds}\frac{\max\left(L_{ds}, L_{ss}\right)}{L_{ds}}, \quad ss_i = n_{ss}\frac{\max\left(L_{ds}, L_{ss}\right)}{L_{ss}}$$

where is the structure score, ds_i and ss_i are the normalized read coverages, and are the total covered length by mapped dsRNA-seq and ssRNA-seq reads, respectively. The total coverage length was used as the normalization constant instead of the total number of mapped reads used

previously, because we believe it is a more reasonable assumption for the transcriptome to have comparable levels of paired/unpaired regions. It is of note that we used a generalized log ratio (glog) instead of normal log-odds because it can tolerate 0 values (positions with no dsRNA or ssRNA read coverage) as well as being asymptotically equivalent to the standard log ratio when the coverage values are large. Only sense-mapping reads were used, as we are entirely concerned with the intra-molecular interactions contributing to the self-folding secondary structure.

A.3.8 Secondary structure and PPS density at upstream Open Reading Frames (uORFs)

Annotated *Arabidopsis* uORFs of high confidence (defined as a purine at the -3 position and a guanine at the +4 position) were extracted from a previously annotated dataset (von Arnim et al., 2014). We then calculated average structure score (see above) and PPS density (average number of PPS covered bases) for uORFs with >10 mapped reads in the regions 50 bp up- or downstream of uORF start codons.

A.3.9 PPS profiles across canonical start codons for transcripts localized to specific cellular compartments

Transcripts were subdivided based on their TAIR10 annotated cellular component gene ontology (mitochondria: 0005739, chloroplast: 0009507, ER: 0005829, Golgi apparatus: 0005794, nucleus: 0005634). PPS density was then calculated and graphed for 50 nt up- and downstream of the start codon as previously described (Silverman et al., 2014).

A.3.10 Structure profile at dsRNase- and ssRNase-identified PPSs

All exonic PPSs and equal sized flanking regions were taken and subdivided into one hundred equal sized bins. The calculated structure scores were averaged for each bin, and the resulting profiles were graphed.

A.3.11 Analysis of alternatively spliced exons and introns

In order to identify specific subsets of alternative splicing events, we took all TAIR10 annotated mRNA transcripts and used the ASTALAVISTA suite (parameters -t asta –i) to identify every annotated alternative splicing event (Foissac and Sammeth, 2007; Sammeth et al., 2008). We then used the ASTALAVISTA code assigned to each event to identify single cassette exons or intron retention sites (0,1^2- or 0,1-2^, respectively). Additionally, we extracted all cassette exon and intron retention events, regardless of adjacent exons, using the list of alternative events and corresponding ASTALAVISTA codes previously described in *Arabidopsis* (Marquez et al., 2012). Taking these annotated events, we then identified the splice donor and acceptor sites of the nearest constitutive introns for our analysis (e.g. if exons 4, 5, and 6 are alternatively spliced together we looked at the donor and acceptor sites at exons 3 and 7, respectively). PPS and structure score profiles were then calculated for regions where the donor exon was ≥ 50 nt, acceptor exon was ≥ 50 nt, and intron was ≥ 60 nt and at least 5 reads mapped to the intron. Thus, these profiles can cover the 50 exonic and 30 intronic nucleotides flanking the splice donor and acceptor sites. P-values were calculated by non-pairwise Wilcoxon tests.

A.3.12 Analysis of alternative polyadenylation sites

We extracted the cleavage and polyadenylation sites previously identified by direct RNA sequencing (Sherstnev et al., 2012) and filtered out sites that were not located within TAIR10 annotated 3’ UTRs. A second filtering step was performed to remove alternative polyadenylation (APA) sites within 60 nt of one another, preventing any overlap between analyzed flanking regions. PPS density and structure score profiles were then calculated for 30 nt flanking each side of these cleavage and polyadenylation sites. P-values were calculated by non-pairwise Wilcoxon tests.

A.3.13 Secondary structure at RBP binding sites

MEME (Bailey et al., 2009) and HOMER (Heinz et al., 2010) were used to identify enriched RBP interaction motifs with parameters -p 8 -dna -nmotifs 100 -maxw 20 -evt 0.01 -maxsize 100000000, and -rna -size given -p 2 respectively. Motifs from **Figures 2.11B-E** were mapped to the genome using HOMER (Heinz et al.) to identify every occurrence of the motifs in pre-mRNAs. We then identified protein-bound and unbound occurrences using our mapped PPSs. Average structure scores for each position were calculated.

A.3.14 Motif and co-occurrence analysis

Motif co-occurrence was defined at the transcript level, and k-means clustering of the resultant weighted adjacency matrix was used to identify clusters of co-occuring motifs. We set k=3 based on manual inspection of clusters on a multidimensional scaling (MDS) plot of the adjacency matrix. Gene Ontology (GO) analysis on the lists of transcripts that contained at least three of the motifs in each cluster was performed using agriGO (Du et al., 2010).

A.3.15 Structure score profile analysis of mRNAs

The structure score for every base of each detected transcript was first calculated using all mapped and spliced reads. In addition to the minimum dsRNA-seq plus ssRNA-seq read coverage requirement discussed above, we only considered mRNAs with intact CDS regions, ≥ 45 nt 5' UTRs, ≥ 140 nt 3' UTRs and a minimum coverage of 50 reads across the entire transcript. To generate profiles, the Z-score of the structure score was calculated for each nucleotide with respect to the graphed window as previously described (Berkowitz et al., 2016).

To analyze profiles across detectable lncRNAs, we divided the length of the transcript into 100 equally sized bins. Taking the average scaled structure score across each bin, we then graphed the profile of these scores.

A.3.16 PPS profile analysis of mRNAs

PPS occupancy was converted to a score at each nucleotide, with a 1 indicating that a protein was bound and a 0 indicating that the nucleotide was unbound. The average PPS occupancy was calculated for all transcripts passing the expression criteria described above. PPS density was then graphed such that the region of highest occupancy was normalized to a density of 1.0.

A.3.17 Mass Spectrometry Analyses

Tryptic peptide extracts were analyzed using nLC-MS/MS (Dionex/LCPackings Ultimate nano-LC coupled to a Thermo LCQ Deca XP+ ion trap mass spectrometer) in duplicate. 1 μl of the peptide sample (in LC buffer A, 0.1% HCOOH (v/v) in 5:95 CH_3CN:H_2O) was first loaded onto a μ-Precolumn (PepMapTM C18, LC-Packings), washed for 4 minutes at a flow rate of 25 μl/min with LC buffer A, then transferred onto an analytical C18-nanocapillary HPLC column (PepMapAcclaim100). Peptides were eluted at 280 nl/min flow rate with a 120 minute gradient of LC buffers A and B (0.1% (v/v) formic acid in 80:20 acetonitrile:water) ranging from 5%-95% B. A fused silica emitter tip with 8-μm aperture (FS360-75-8-N-5-C12; New Objective) mounted to a Thermo nanospray ionization (NSI) source at 1.8 kV was used for positive ionization of peptides. Mass spectra were collected using Thermo Xcalibur 2.0 software. The top 3 principal ions from each MS scan were trapped and fragmented during the chromatographic gradient, using dynamic exclusion to maximize detection of ions (range 200-2000 m/z). The trapped ions were subjected to collision-induced dissociation (CID) with He, and ~4000 spectra (MS/MS) were collected to cover the entire chromatography elution profile.

A.3.18 Spectral Data Analyses and Protein ID

Experimentally collected MS/MS tandem data were searched against the *Arabidopsis* Proteome Database (NCBI, latest version) using Thermo Proteome Discoverer 1.4 software. The search was restricted to full trypsin digestion with a maximum of 3 missed cleavages and

potential modifications for methionine (oxidation) and cysteine (carbamidomethylation); other parameters were standard for LCQ Deca XP+ instrumentation. Peptide filters were set to standard Xcorr vs charge state values; Xcorr = (1.5, 2.0, 2.25, 2.5) for charges (+1,+2,+3,+4), respectively. Spectral assignments were manually scrutinized to validate the reliability of the protein identifications.

A.3.19 HAMR analysis of mRNA-seq data

To perform HAMR analysis we first isolated all uniquely mapping sequencing reads, removing multi-mapping reads from our analysis. Using the Picard analysis software, we added a read group to the bam file, and then split the cigar string using the GATK software. This step divides junction spanning reads into individual lines in the bam file, therefore when HAMR is run the nucleotides across the junction will not be erroneously called as mismatches in the intron. We then ran the HAMR pipeline, filtering out the first and last nucleotides of each read, which are enriched for adapter sequences, and any reads with a quality score lower than Q30. Taking these high quality reads we then search for covalent modifications with an FDR of <0.05. HAMR modifications between biological replicates are then concatenated together with redundant modifications collapsed for subsequent analyses.

A.3.20 Calculated modification distribution across a metagene

To analyze modification localization across detectable mRNAs, we divided the length of the 5' UTR, CDS, and 3' UTR into 100 equally sized bins each. Taking the average number of modifications across each bin, we then graphed the profile of these scores. To account for varying coverage, we only examined nucleotides that are HAMR accessible in all three graphed samples. Therefore, any nucleotide that did not have a sequencing depth of at least 50 reads in the nuclear, cytoplasmic, and whole tissue samples, was omitted from the analysis.

A.3.21 Calculating proportion uncapped

The proportion uncapped or measure of transcript instability was calculated as the log ratio of the reads per million (RPM) of an annotated mRNA in the GMUCT library and dividing it by the RPM of that mRNA in the whole tissue mRNA-seq library. Therefore, a higher proportion uncapped indicates more reads in the GMUCT library and therefore a lower stability for the mRNA.

A.4 DATA AND SOFTWARE AVAILABILITY

A.4.1 Chapter 1 Accession Numbers

The raw and processed data for PIP-seq from our analyses have been deposited into the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE58974.

A.4.2 Chapter 2 Accession Numbers

The raw and processed data for PIP-seq from our analyses have been deposited into the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE86459.

A.4.3 Chapter 1 Genome browser view

The sequencing data presented here is also available through JBrowse genome browser: http://gregorylab.bio.upenn.edu/jbrowse/?data=data/At_total_nuc.

A.4.4 Chapter 2 Genome browser view

The sequencing data presented here is also available through the EPIC-CoGe genome browser (Lyons and Freeling, 2008): https://genomevolution.org/coge/NotebookView.pl?nid=1767.

APPENDIX B: PROTEIN INTERACTION PROFILE SEUQENCING (PIP-SEQ) PROTOCOL

This section refers to work from:

- **Foley, S.W.** and Gregory, B.D. (2016). Protein Interaction Profile Sequencing (PIP-seq). *Current Protocols in Molecular Biology*. *116:27.5.1-27.5.15.* PMID: 27723083

B.1 INTRODUCTION

Many biological processes are dependent on proper post-transcriptional regulation (Braunschweig et al., 2013; Ebert et al., 2012). All of these post-transcriptional processes require a specific collection of RNA binding proteins (RBPs). The interaction of RBPs with their target transcripts is regulated by both the primary sequence and secondary structure, or folding pattern, of the RNA molecule. Previous studies have primarily probed RNA-protein interactions via protein-centric approaches. Specifically, these studies utilized RBP-specific immunoprecipitation, followed by high-throughput sequencing or microarray analyses on the co-purified RNA sequences (Hafner et al., 2010; Licatalosi et al., 2008; Ule et al., 2003). While many insights have been made by these studies, they are limited to uncovering only the binding sites of a single protein.

Protein interaction profile sequencing (PIP-seq) is a genome-wide approach to simultaneously identify protein-bound RNA sequences as well as RNA secondary structure (Gosai et al., 2015; Silverman et al., 2014). PIP-seq utilizes a ribonuclease (RNase)-mediated protein footprinting approach to identify RBP-bound RNA sequences and RNA secondary structure on a global scale. Briefly, whole tissue or tissue culture cells are first treated with formaldehyde in order to crosslink RNA-protein interactions (Basic Protocol 1). The crosslinked material is then lysed and separated into four samples, two footprinting samples, and two structure only samples. In the structure only samples, proteins are first denatured and digested by treatment with SDS and Proteinase K. These samples are then treated with an RNase that specifically digests single-stranded RNA (ssRNase) or double-stranded RNA (dsRNase).

Conversely, the footprinting samples are first treated with the structure-specific RNases prior to protein denaturation and digestion followed by RNA extraction (Basic Protocol 2). Lastly, the four samples undergo strand-specific high-throughput sequencing library preparation (Basic Protocol 3; **Figure B.1**). After sequencing, RNA secondary structure can be inferred by comparing the coverage of each structure only sample. Additionally, protein-bound sequences can be identified by searching for sequences enriched in the footprinting samples as compared to the structure only samples. Together, these data produce a simultaneous view of the global landscapes of both RNA secondary structure and RNA-protein interactions.

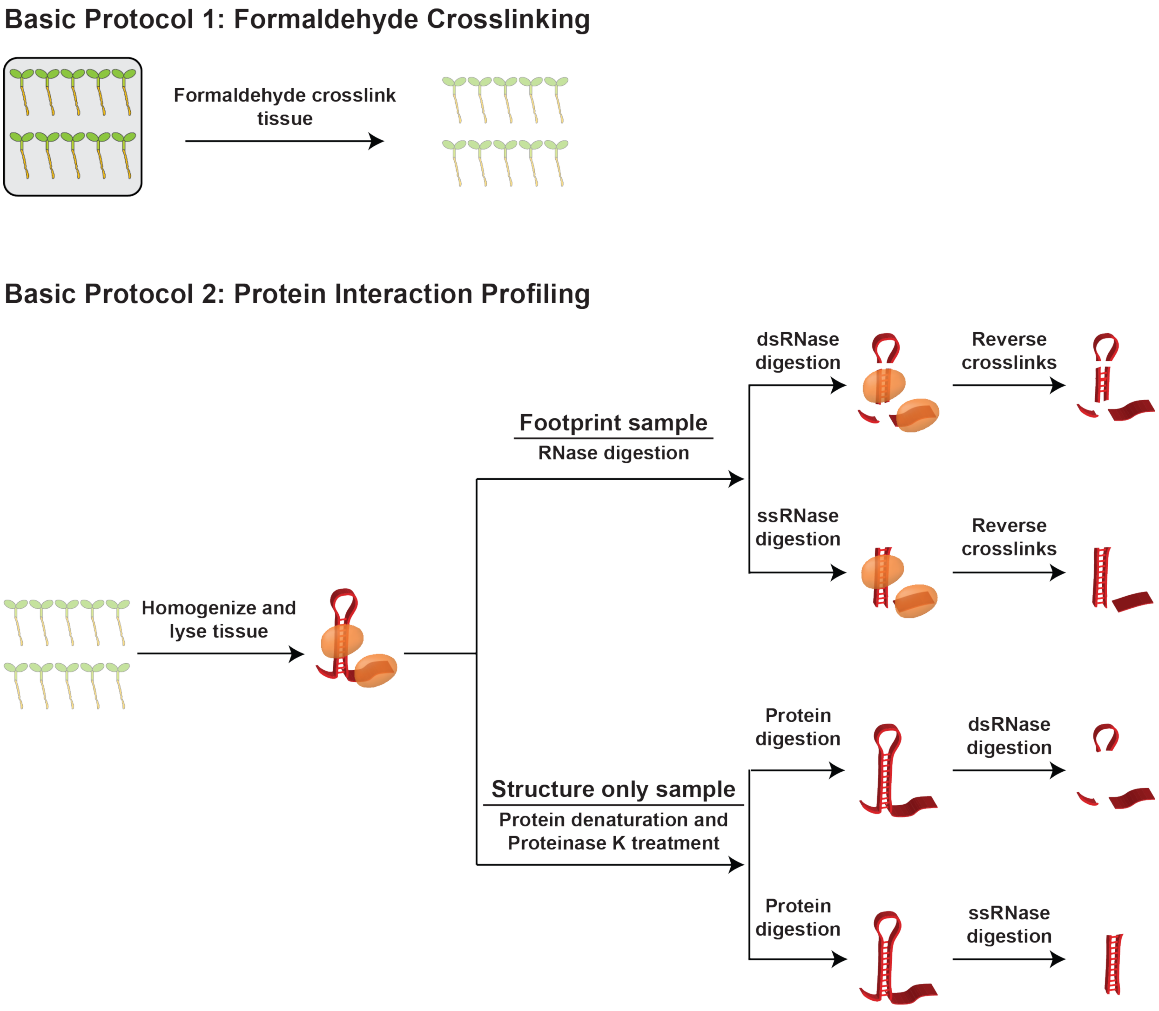


Figure B.1: Overview of PIP-seq protocol
This diagram illustrates all three basic protocols, including formaldehyde crosslinking (Basic Protocol 1), protein interaction profiling (Basic Protocol 2), and strand-specific RNA library preparation (Basic Protocol 3). Although seedlings are illustrated as an example, PIP-seq can be used with a variety of tissue types.

B.2 REAGENTS AND SOLUTIONS

B.2.1 Crosslinking buffer

810 μL 37% Formaldehyde (1% final) (Sigma, 252549)

(optional) 6 μL SilWet L-77 (0.02% final) (Lehle Seeds, VIS-30)

Distilled, deionized H_2O, up to 30 mL

Prepare fresh immediately before use

B.2.2 RIP Buffer

12.5 mL 1 M Tris-HCl, pH = 8.6 (25 mM final)

37.5 mL 2 M KCl (150 mM final)

5 mL 0.5 M EDTA pH = 8.0 (5 mM final) (Fisher Scientific, BP2483)

2.5 mL Igepal CA-630 (0.5% final) (Sigma, I8896)

Distilled, deionized H_2O, up to 500 mL

Store at 4ºC up to one year

Prepare 10 mL fresh immediately before use by adding:

1 tablet Protease Inhibitor Mini Cocktail (Roche, 11836170001)

5 μL RNaseOUT (Life Technologies, 10777019)

50 μL 0.1M DTT (Life Technologies, 18064-014)

B.2.3 16x RNase Stop Buffer

5 mL 20% SDS (10% final) (Denville Sciences, CS5585-28)

100 μL 0.5 M EDTA pH = 8.0 (5mM final) (Fisher Scientific, BP2483)

1 μL 20 mg/mL Proteinase K (Denville Sciences, CB3210-5)

Distilled, deionized H_2O, up to 10 mL

Prepare fresh immediately before use

B.2.4 1x DNase solution

550 μL DEPC-treated H_2O

3.85 mL RDD buffer (Qiagen, 79254)

Use solution to resuspend lyophilized RNase free DNase (Qiagen, 79254)

Prepare 80 μL aliquots and store at -20ºC

B.2.5 DSN Hybridization Buffer

200 μL 1 M HEPES buffer solution (200 mM final) (Life Technologies, 15630)

400 μL 5 M NaCl (2 M final) (Life Technologies, AM9937)

400 μL Nuclease free H_2O

Aliquot and store at -20ºC

B.2.6 DSN STOP Buffer

10 mM EDTA

B.3 BASIC PROTOCOL 1: FORMALDEHYDE CROSSLINKING OF TISSUE

In order to identify protein-bound RNA sequences, the RNA-protein interactions must be preserved after cell lysis. To accomplish this, the tissue is treated with a 1% formaldehyde solution under a vacuum to crosslink these *in vivo* interactions. The crosslinking reaction is then quenched by vacuum infiltrating 125 mM glycine into the tissue, followed by washing with distilled, deionized water, prior to performing protein interaction profiling.

B.3.1 Materials

Tissue sample

Crosslinking buffer

1 M Glycine (Sigma, G8790)

Liquid N_2

Scale

50 mL conical tubes

End-over-end rotator

Bell-shaped vacuum container

Vacuum line

Paper towels

B.3.2 Protocol

1. Transfer the tissue to a 50 mL conical tube containing 30 mL of crosslinking buffer and let rotate end-over-end for 1 min.

    We have previously used 1-3 grams of plant tissue in 30 mL of crosslinking buffer. Be sure that the buffer is in excess, and the formaldehyde solution can properly penetrate the tissue.

    We have used SilWet L-77 for whole seedling and leaf tissue to increase permeability.

2. Remove the lid and place the conical tube in the bell-shaped vacuum container connected to a vacuum line, and let incubate under vacuum for 10 min.
3. Quench the reaction by adding 4.25 mL 1 M Glycine to a final concentration of 125 mM.
4. Mix the quenched solution thoroughly by inversion, remove the cap, and place under vacuum for 5 min.
5. Wash the tissue 5 times with distilled, deionized water.
6. Pat the tissue dry with a paper towel, and snap freeze in liquid N_2.
7. Store the tissue at -80ºC for up to six months.

B.4 BASIC PROTOCOL 2: PROTEIN INTERACTION PROFILING

Formaldehyde crosslinked tissue (Basic Protocol 1) is then homogenized, lysed, and treated with both RNases and proteases during protein interaction profiling. The two structure only libraries first have proteins denatured and digested with an SDS and Proteinase K solution, followed by digestion with either an ssRNase or dsRNase. The coverage across these two libraries will later be compared to infer secondary structure. Additionally, two footprinting samples will be prepared. These samples first undergo RNase digestion prior to protease treatment. These proteins will inhibit RNA digestion creating protein protected sites (PPSs), sequences that are enriched in the footprinting sample when compared to the structure only samples.

B.4.1 Materials

Formaldehyde crosslinked tissue (Basic Protocol 1)

Liquid N_2

RIP buffer

DNase solution

16x Stopping Solution

3M NaOAc (pH=5.5) (Life Technologies, AM9740)

100% EtOH (Decon Labs, 2716)

70% EtOH

Qiazol (Qiagen, 79306)

RNaseONE buffer (Promega, M4261)

RNaseONE (Promega, M4261)

10x RNA Structure Buffer (Life Technologies, AM2275)

RNase V1 (Life Technologies, AM2275)

Mortar and pestle

1.7 mL tubes

Plastic pestle

2.0 mL tubes

Centrifuge

37ºC heat block

65ºC heat block

B.4.2 Protocol

1. Homogenize the tissue with a mortar and pestle.
2. Resuspend tissue in 850 μL RIP buffer and transfer to pre-cooled 1.7 mL tube

3. Grind with plastic pestle, then pass through 1000 μL and 200 μL pipette tips ~20 times each.
4. Add 160 μL DNase solution and incubate at room temperature (RT) for 30 min.
5. Aliquot 250 μL of lysate and 600 μL RIP into two 2.0 mL tubes for structure only library samples and two 1.7 mL tubes for footprinting library samples.

To prepare ssRNase-treated structure only library

6. To one of the 850 μL samples in 2.0 mL tubes, add 75 μL 16x stopping solution and incubate at RT for 15 min.
7. Precipitate RNA by adding 90 μL 3M NaOAc and 1 mL 100% EtOH, then store at -80ºC for more than 1 hour.
8. Centrifuge at 4ºC for 45 min at 20,000 x *g*.
9. Remove supernatant and add 700 μL 70% EtOH to wash.
5. Centrifuge 4ºC for 5 min at 20,000 x *g*.
6. Remove supernatant and let pellet air dry for 10 min.
7. Resuspend in 850 μL RIP buffer.

    The pellet will be difficult to resuspend, pipet and vortex thoroughly.

8. Add 100 μL RNase ONE buffer and 10 μL RNase ONE.
9. Incubate at 37ºC for 1 hour 15 min, mixing ever 15 min.
10. Add 75 μL 16x RNase Stop buffer.
11. Incubate at RT for 10 min.
12. Transfer samples to 65ºC for 2 hours mixing every 30 minutes.
13. Divide the sample into two 2.0 mL tubes with ~500 μL each.
14. Add 700 μL QIAzol to each tube, vortex thoroughly, and store at -20ºC.

    The samples can remain at -20ºC overnight.

To prepare dsRNase-treated structure only library

15. To the other 850 μL sample in 2.0 mL tubes, add 75 μL 16x RNase Stop buffer and incubate at RT for 15 min.
16. Precipitate RNA by adding 90 μL 3M NaOAc, 1 mL 100% EtOH and store at -80ºC for more than 1 hour.
17. Centrifuge at 4ºC for 45 min at 20,000 x *g*.
18. Remove supernatant and add 700 μL 70% EtOH to wash.
19. Centrifuge at 4ºC for 5 min at 20,000 x *g*.
20. Remove supernatant and let pellet air dry for 10 min.
21. Resuspend in 850 μL RIP buffer.

    The pellet will be difficult to resuspend, pipet and vortex thoroughly.
22. Add 100 μL 10x RNA Structure Buffer and 25 μL RNase V1.
23. Incubate at RT for 1 hour 15 min, mixing ever 15 min.
24. Add 75 μL 16x RNase Stop buffer.
25. Incubate at RT 10 minutes.
26. Transfer samples to 65ºC for 2 hours mixing every 30 minutes.
27. Divide the sample into two 2.0 mL tubes with ~500 μL each.
28. Add 700 μL QIAzol to each tube, vortex thoroughly, and store at -20ºC.

    The samples can remain at -20ºC overnight.

To prepare ssRNase-treated footprinting library

29. To one of the 850 μL samples in 1.7 mL tubes, add 100 μL RNase ONE buffer and 10 μL RNase ONE.
30. Incubate at 37ºC for 1 hour, mixing ever 15 min.
31. Add 75 μL 16x RNase Stop buffer.
32. Incubate at RT 10 minutes.

33. Transfer samples to 65ºC for 2 hours mixing every 30 minutes.
34. Divide the sample into two 2.0 mL tubes with ~500 μL each.
35. Add 700 μL QIAzol to each tube, vortex thoroughly, and store at -20ºC.

    The samples can remain at -20ºC overnight.

To prepare dsRNase treated structure only library

36. To the other 850 μL sample in 1.7 mL tubes, add 100 μL 10x RNA Structure Buffer and 25 μL RNase V1.
37. Incubate at RT for 1 hour, mixing ever 15 min.
38. Add 75 μL 16x RNase Stop buffer.
39. Incubate at RT 10 minutes.
40. Transfer samples to 65ºC for 2 hours mixing every 30 minutes.
41. Divide the sample into two 2.0 mL tubes with ~500 μL each.
42. Add 700 μL QIAzol to each tube, vortex thoroughly, and store at -20ºC.

    The samples can remain at -20ºC overnight.
43. Perform Qiagen miRNeasy RNA extraction on all samples following the manufacturer's protocol.

B.5 BASIC PROTOCOL 3: STRAND-SPECIFIC HIGH-THROUGHPUT SEQUENCING LIBRARY PREPARATION

This protocol describes the process of producing a strand-specific high-throughput sequencing library using the RNA isolated at the end of Basic Protocol 2. This is a modified version of the Illumina strand-specific high-throughput library preparation protocol, which includes several size selection steps that are necessary for removal of adapter duplexes. In order to remove highly abundant RNA sequences such as rRNA and tRNA this protocol utilizes a Duplex Specific Nuclease (DSN) treatment. This treatment first denatures the cloned DNA at 98ºC, and then allows the DNA molecules to slowly reanneal for 5 hours at 68ºC. The high temperature

leads to slow reannealing, resulting in the most highly abundant transcripts reannealing more quickly than those that are less abundant. The double-stranded DNA sequences are then digested by the by the DSN, dramatically reducing the rRNA and tRNA content.

B.5.1 Materials

Protein interaction profile RNA (Basic Protocol 2)

10x Fragmentation Reagent (Life Technologies, AM8740)

Fragmentation Stop Solution (Life Technologies, AM8740)

DEPC-treated H_2O

5 mg/mL Glycogen, ultrapure (Life Technologies, AM9510)

3 M NaOAc (pH = 5.5) (Life Technologies, AM9740)

100% EtOH (Decon Labs, 2716)

80% EtOH

T4 DNA ligase buffer (NEB biolabs, B0202S)

T4 Polynucleotide kinase (NEB biolabs, M0201S)

10 mM ATP (Life Technologies, AM8110G)

10x TBE (Bio-Rad, 161-0733)

Gel loading buffer II (Life Technologies, AM8546G)

10 bp DNA ladder (Life Technologies, 10821-015)

10 mg/mL Ethidium Bromide (Life Technologies, 15585-011)

0.3 M NaCl

5 μM 3' Adapter (RA3) (TGGAATTCTCGGGTGCCAAGG)

RNA Ligase Buffer (NEB, B0216L)

RNaseOUT (Life Technologies, 10777019)

200 U/μL Epicenter T4 RNA Ligase 2, truncated (NEB, M0242S)

25 μM 5' Adapter (RA5) (GUUCAGAGUUCUACAGUCCGACGAUC)

T4 RNA Ligase 1 (NEB, M0204S)

RNA RT Primer (RTP) (GCCTTGGCACCCGAGAATTCCA)

5x First Strand Buffer (Life Technologies, 18064-014)

50 mM dNTPs

100 mM DTT (Life Technologies, 18064-014)

SuperScript II Reverse Transcriptase (Life Technologies, 18064-014)

2x Phusion Mix (NEB, M0531S)

5 mM Betaine (MP Biomedicals, 215046180)

10 μM RNA PCR Primer (RTP) (GCCTTGGCACCCGAGAATTCCA)

10 μM RNA PCR Primer Index

25 bp DNA ladder (Life Technologies, 10597-011)

Hybridization buffer

10x DSN Master Mix (Evrogen, EA001)

DSN Enzyme (Evrogen, EA001)

DSN STOP solution (Evrogen, EA001)

1.7 mL tubes

70ºC heat block

37ºC heat block

15% TBE-Urea gels 1.0 mm, 10 well (Life Technologies, EC6885BOX)

Gel box for running pre-poured gels (Life, Technologies, EI0001)

18-gauge needles

Razor blades

Gel Breaker Tubes (IST Engineering, 388-100)

2.0 mL tubes

Spin-X columns (Costar, 8160)

200 μL PCR tubes

Thermocycler

6% TBE gels 1.0 mm, 10 well (Invitrogen, EC6265BOX)

B.5.2 Protocol

RNA Fragmentation:

1. Add 1 μL of 10x Fragmentation Reagent to 9 μL RNA.
2. Incubate at 70ºC for 3 min.
3. **Immediately** add 1 μL Stop Solution
4. Bring volume up to 100 μL with DEPC-treated H_2O.
5. Precipitate RNA by adding 3 μL Glycogen, 1/10 volume 3M NaOAc (pH=5.5), and 3x volume 100% EtOH.
6. Store at -80ºC for at least 2 hours.
7. Centrifuge samples at 4ºC for 80 min at 20,000 x *g*.
8. Wash samples with 700 μL 80% EtOH.
9. Remove supernatant and let samples air dry for 10 min.
10. Resuspend in 16 μL DEPC-treated H_2O.
11. Incubate on ice for 25 min.

T4 Polynucleotide Kinase Treatment

12. Add 2 μL NEB T4 DNA ligase buffer, 1 μL T4 Polynucleotide kinase, 1 μL 10 mM ATP to 16 μL RNA.
13. Incubate at 37ºC for 1 hour.
14. Add 80 μL DEPC treated H_2O to bring volume to 100 μL.
15. Precipitate RNA (Basic Protocol 3 Steps 5-9)
16. Resuspend precipitant in 10 μL DEPC-treated H_2O.
17. Incubate on ice for 25 min.

Size Selection #1

18. Assemble 15% TBE-Urea polyacrylamide gel in a gel box with ~500 mL 1x TBE.
19. Pre-clear wells with 18-gauge needle.
20. Pre-run gel at 155 V for 25 min.
21. Add 10 μL Gel Loading Buffer II to RNA samples.
22. Heat samples at 70ºC for 5 min, then snap cool on ice for 3 min.
23. Prepare 1.5 μg 10 base pair (bp) DNA ladder in 10 μL DEPC-treated H2O and 10 μL Gel Loading Buffer II
24. Pre-clear wells again with 18-gauge needle.
25. Load samples into gel and run at 155 V for 90 min or until bromphenol blue has traveled 80% of the gel.
26. Stain gel in 100 mL 1x TBE and 14 μL 10 mg/mL Ethidium Bromide for 10 min.
27. Using an ethanol-cleaned razor blade, excise gel region from 15-150 nucleotides for each sample on the gel.

    Make sure to use clean razor blades for each sample to avoid contamination.
28. Place gel slices in Gel Breaker Tube inside of 2.0 mL tube, and centrifuge samples at RT for 2 min at 20,000 x *g*.
29. Add 300 μL 0.3 M NaCl and rotate end over end for 4 hours.
30. Transfer the entire sample to a Spin-X column and centrifuge samples at RT for 2 min at 20,000 x *g*.
31. Transfer flowthrough to a new tube and precipitate RNA (Basic Protocol 3 Steps 5-9)
32. Resuspend RNA in 5 μL DEPC-treated H_2O
33. Allow samples to resuspend on ice for 25 min.

Ligate 3' and 5' Adapters

34. Transfer the 5 μL RNA samples to a 200 μL PCR tube and add 1 μL 5 μM 3' Adapter (RA3).

35. Mix thoroughly, and incubate at 70ºC for 2 min, then 4ºC for 2 min.
36. Add 4 μL of the following ligation reagents and mix thoroughly.

    2 μL RNA Ligase Buffer

    1 μL RNaseOUT

    1 μL 200 U/μL Epicenter T4 RNA Ligase 2, truncated
37. Incubate at 28ºC 1 hour 15 min,
38. Near the end of the incubation, transfer 1 μL 25 μM 5' adapter (RA5) to 70ºC for 2 min and snap cool on ice 2 min.

    The total amount of 5' adapter needed for all ligations can be denatured in a single tube
39. Add 1 μL 10 mM ATP and 1 μL T4 RNA Ligase 1 to the tube containing the cooled 5' adapter.

    This can be made into a master mix with the denatured 5' adapter for all samples to avoid pipetting errors.
40. Aliquot 3 μL of the 5' adapter mix to each RNA sample and incubate at 28ºC for 1 hour.

    Store samples at -20ºC, or proceed to Size Selection #2.

Size Selection #2

41. Prepare 15% TBE-Urea polyacrylamide gel (Basic Protocol 3 Steps 18-20)
42. Add 10 μL Gel Loading Buffer II to RNA ligase reaction and incubate at 70ºC for 5 min, then snap cool on ice for 3 min.
43. Run the samples on the 15% TBE-Urea polyacrylamide gel and stain with ethidium bromide (Basic Protocol 3 Steps 23-26)
44. Using an ethanol-cleaned razor for each sample on the gel, excise gel from 65-200 nucleotides.

    Make sure to use clean razor blades for each sample to avoid contamination.
45. Elute and precipitate RNA from gel slice (Basic Protocol 3 Steps 27-31).

46. Resuspend RNA in 6 μL DEPC-treated H_2O.
47. Incubate on ice for 25 min.

Reverse Transcription

48. Preheat a thermocycler to 70ºC.
49. Transfer 6 μL of the RNA samples to separate wells in a strip 200 μL PCR tubes.
50. Add 1 μL 100 μM RT Primer (RTP) and mix thoroughly.
51. Incubate samples at 70ºC for 2 min, then 4ºC for 2 min.
52. Add 5.5 μL of the following RT reagents and mix thoroughly.

    2 μL 5x First Strand Buffer

    0.5 μL 50mM dNTP mix

    1 μL 100 mM DTT

    1 μL RNaseOUT

    1 μL Superscript II Reverse Transcriptase
53. Incubate the RT reaction at 50ºC for 1 hour.
54. Proceed directly to PCR Amplification.

PCR Amplification

55. Make PCR master mix

    50 μL Phusion Mix 2x

    33.5 μL 5 mM Betaine

    2 μL RNA PCR Primer (RP1)
56. Add 85.5 μL PCR master mix to RT reaction
57. Add 2 μL of RNA PCR Primer Index to each sample.

    Each sample should have a different RNA PCR Primer Index allowing multiplexing during sequencing.

58. Divide each 100 μL sample between four 200 μL PCR tubes, ~25 μL per tube.
59. Incubate samples on thermocycler:

1 cycle:	98ºC 30 seconds
12 cycles:	98ºC 10 seconds
	60ºC 30 seconds
	72ºC 15 seconds
1 cycle:	72ºC 10 min
	Hold at 4ºC

60. Recombine the four PCR tubes.
61. Precipitate DNA by adding 3 μL Glycogen, 1/10 volume 3M NaOAc (pH=5.5), and 3x volume of 100% EtOH.
62. Store at -80ºC for at least 2 hours.
63. Centrifuge samples at 4ºC for 45 min at 20,000 x *g*.
64. Wash samples with 700 μL 70% EtOH.
65. Centrifuge samples at 4ºC for 5 min at 20,000 x *g*.
66. Remove supernatant and let samples air dry for 10 min.
67. Resuspend in 10 μL Nuclease-free H_2O.
68. Incubate on ice for 25 min.

Size Selection #3

69. Assemble 6% TBE-Urea polyacrylamide gel in a gel box with ~500 mL 1x TBE.
70. Add 10 μL Gel Loading Buffer II to DNA samples.
71. Prepare 1.5 μg 25 bp DNA ladder in 8.5 μL DEPC treated H2O and 10 μL Gel Loading Buffer II
72. Pre-clear wells with 18-gauge needle.
73. Load samples into gel and run at 155 V for 30 min or until bromphenol blue has traveled 80% of the gel.

74. Stain gel in 100 mL 1x TBE and 14 μL 10 mg/mL Ethidium Bromide for 10 min.
75. Using an ethanol-cleaned razor for each sample on the gel, excise gel from 135 bp to the end of the signal, being carful to avoid the adapter-adapter band at 120 bp.

    Make sure to use clean razor blades for each sample to avoid contamination.
76. Place gel slices in Gel Breaker Tube inside of 2.0 mL tube, and centrifuge samples at RT for 2 min at 20,000 x *g*.
77. Add 300 μL 1x NEB Buffer 2 and rotate end over end for 2 hours.
78. Transfer the entire sample to a Spin-X column and centrifuge samples at RT for 2 min at 20,000 x *g*.
79. Transfer flowthrough to a new tube and precipitate DNA (Basic Protocol 3 Steps 61-66)
80. Resuspend the precipitated DNA in 15.5 μL Nuclease Free H_2O
81. Allow samples to resuspend on ice for 25 min.

Duplex Specific Nuclease (DSN) Treatment

82. Put a total of 100 ng of sample DNA into 13.5 μL Nuclease-free H_2O in 200 μL PCR tube.

    Store the remaining library at -80ºC. The DSN treatment can be repeated if more library is necessary for sequencing.

    Although 100 ng is ideal, we have generated successful libraries with as few as 20 ng starting material. Be sure to use a consistent amount for all PIP-seq samples.
83. Add 4.5 μL DSN Hybridization buffer, and mix thoroughly.
84. Incubate samples in the thermocycler:

    98ºC 2 min

    68ºC 5 hours

    The 5 hour incubation has been found to reduce rRNA and tRNA in Arabidopsis. For other species a range of incubation times should be tested.

85. After 4.5 hours dilute 4 μL 10x DSN Master buffer in 16 μL Nuclease-free H_2O to make 2x DSN Master buffer.
86. Incubate 2x DSN Master buffer at 68ºC for the remainder of the 5 hour incubation.
87. Add 20 μL 2x DSN Master buffer to each sample, and mix thoroughly.

    Do not allow samples to cool down*, mix by pipetting and* ***briefly*** *centrifuge.*
88. Incubate samples at 68ºC for another 10 min.
89. Add 2 μL DSN enzyme to DNA samples, and mix thoroughly.

    Do not allow samples to cool down*, mix by pipetting and* ***briefly*** *centrifuge.*
90. Incubate samples at 68ºC for another 25 min.
91. Add 40 μL DSN STOP solution, and mix thoroughly.

    Do not allow samples to cool down*, mix by pipetting*
92. Allow to cool at 4ºC for 2 min.
93. Transfer samples to 1.7 mL tubes and add 20 μL Nuclease-free H_2O, bringing the volume up to 100 μL.
94. Precipitate DNA (Basic Protocol 3 Steps 61-66)
95. Resuspend DNA in 6 μL Nuclease Free H_2O
96. Allow samples to resuspend on ice for 25 min.

PCR Amplification #2

97. Make PCR master mix

    50 μL Phusion Mix 2x

    40 μL 5 mM Betaine

    2 μL RNA PCR Primer
98. Add 92 μL PCR master mix to RT reaction
99. Add 2 μL of RNA PCR Primer Index to each sample.

    Be sure that this is the same PCR Primer Index used in Basic Protocol 3 Step 57.

100. Divide each 100 μL sample between four 200 μL PCR tubes, ~25 μL per tube.

101. Incubate samples on thermocycler:

1 cycle:	98ºC 30 seconds
12 cycles:	98ºC 10 seconds
	60ºC 30 seconds
	72ºC 30 seconds
1 cycle:	72ºC 5 min
	Hold at 4ºC

102. Recombine the four PCR tubes.

103. Precipitate DNA (Basic Protocol 3 Steps 61-66)

104. Resuspend in 10 μL Nuclease Free H_2O.

105. Incubate on ice for 25 min.

Size Selection #4

106. Run samples on 6% TBE-Urea polyacrylamide gel and excise bands greater than 135 bp as in Basic Protocol 3 Steps 69-79.

107. Resuspend DNA in 30 μL Nuclease-free H_2O

108. Allow samples to resuspend on ice for 25 min.

You should now have a completed set of PIP-seq libraries for high-throughput sequencing that allow the simultaneous analysis of RNA secondary structure and RNA-protein interactions.

B.6 COMMENTARY

B.6.1 Background Information

The study of post-transcriptional regulation is limited only by the techniques currently available. While the past decade has seen the development of numerous approaches involving single protein immunoprecipitation followed by sequencing (Ule et al., 2003), more recent

protocols have attempted to identify RNA sequences bound globally throughout a sample. In addition to PIP-seq, another approach is global photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation (gPAR-CLIP). This technique utilizes the ribonucleoside 4-thiouridine, which forms covalent bonds with adjacent proteins when exposed to 365 nm UV light. This covalent bond causes misincorporation of guanine by reverse transcriptase, resulting in uracil-to-cytosine mismatches in the sequencing reads at protein-bound sites (Baltz et al., 2012). This technique requires both the uptake and incorporation of 4-thiouridine and UV crosslinking treatment; therefore it is only applicable in a tissue culture system. Alternatively, PIP-seq utilizes a formaldehyde crosslinking step, allowing it to be used on whole tissues.

B.6.2 Critical Parameters and Troubleshooting

The PIP-seq protocol is dependent on complete digestion by structure-specific RNases. Unlike previous protocols that use low levels of RNase V1 to study single-hit kinetics (Wan et al., 2014), PIP-seq fully digests the double- and single-stranded RNA in each corresponding sample. Therefore, the relatively high concentration of these RNases, and the long incubations are necessary to best gauge secondary structure and protein binding. Additionally, the high levels of SDS in the 16x RNase Stop Solution are necessary to fully denature the RBPs in both samples.

After preparing high-throughput sequencing libraries, two common concerns are the presence of unwanted RNA populations (rRNA and tRNA), as well as adapter duplexes and other sequencing artifacts. In order to reduce highly abundant RNA classes this protocol utilizes a DSN library treatment step, and recommends a 5 hour 68°C reannealing time, which has been used for plant tissue samples. The reannealing time must be determined empirically for different samples, which we have found can range from 4 to 9 hours. Following sequencing library completion, a small aliquot of library can be taken and used for TOPO cloning and Sanger sequencing in order to estimate the percentage of clones containing unwanted RNA populations. Additionally, a higher quantity of low quality sequencing reads can be attributed to adapter duplex

contamination that accumulate during the cloning process. Therefore, it is crucial to remove any unwanted bands during the size selection steps of this protocol (**Figure B.2**).

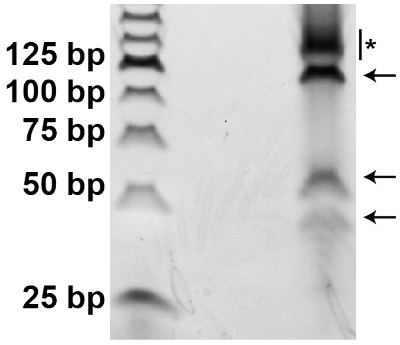


Figure B.2: Size selection gel from strand-specific library preparation
A gel image taken during Size Selection #4. The DSN-treated library post-amplification is indicated by the asterisk (*) and the unincorporated primers and adapter duplexes are indicated by the arrows.

B.6.3 Anticipated Results

Agilent Bioanalyzer 2100 analysis of RNA from protein interaction profiling should reveal RNA fragments from ~15 nt to ~1,000 nt. The footprinting samples will contain larger fragments than the structure only samples (**Figure B.3**). This is due to proteins protecting longer regions of RNA from RNase digestion, and indicates high quality samples ready for library preparation.

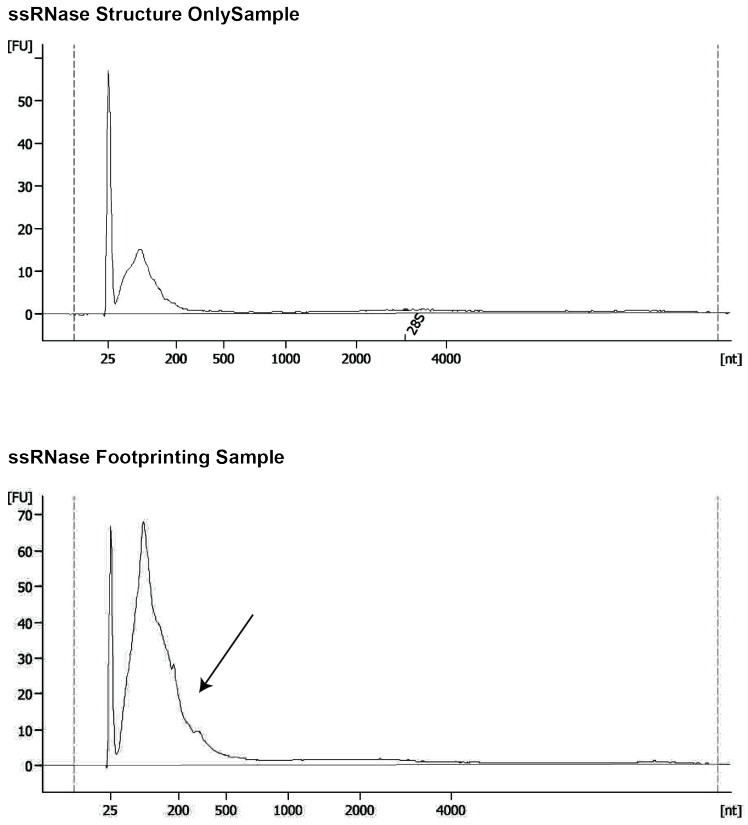


Figure B.3: Agilent Bioanalyzer 2100 analysis of RNA from protein interaction profiling
The bioanalyzer traces from a structure only (top) and a footprinting (bottom) sample. The arrow indicates the longer RNA fragments in the footprinting sample.

B.6.4 Time Considerations

After harvesting tissue the formaldehyde crosslinking, protein interaction profiling, and RNA extraction will take 1-2 days. The formaldehyde crosslinking step takes only ~45 minutes, while protein interaction profiling protocol will last ~11 hours. Although both protocols can all be completed in one day, the samples can be stored at the end of each protocol. The strand-specific library preparation can takes 4-8 days. The samples can be stored in EtOH at -80ºC during each precipitation step. Therefore, every precipitation is a stopping point.

REFERNCES

Alarcón, C.R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S., and Tavazoie, S.F. (2015). HNRNPA2B1 is a mediator of m6A-dependent nuclear RNA processing events. Cell *162*, 1299–1308.

Anantharaman, V., Koonin, E.V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res. *30*, 1427–1464.

Andersen, J., Delihas, N., Hanas, J.S., and Wu, C.W. (1984). 5S RNA structure and interaction with transcription factor A. 1. Ribonuclease probe of the structure of 5S RNA from Xenopus laevis oocytes. Biochemistry (Mosc.) *23*, 5752–5759.

Antal, M., Boros, E., Solymosy, F., and Kiss, T. (2002). Analysis of the structure of human telomerase RNA in vivo. Nucleic Acids Res. *30*, 912–920.

Ares, M., and Igel, A.H. (1990). Lethal and temperature-sensitive mutations and their suppressors identify an essential structural element in U2 small nuclear RNA. Genes Dev. *4*, 2132–2145.

Arnez, J.G., and Steitz, T.A. (1994). Crystal structure of unmodified tRNA(Gln) complexed with glutaminyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. Biochemistry (Mosc.) *33*, 7560–7567.

von Arnim, A.G., Jia, Q., and Vaughn, J.N. (2014). Regulation of plant translation by upstream open reading frames. Plant Sci. *214*, 1–12.

Asakura, Y., and Barkan, A. (2007). A CRM domain protein functions dually in group I and group II intron splicing in land plant chloroplasts. Plant Cell *19*, 3864–3875.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acid Res. *37*, W202–W208.

Baltz, A.G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. (2012). The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. Mol Cell *46*, 674–690.

Bates, T.R., and Lynch, J.P. (1996). Stimulation of root hair elongation in Arabidopsis thaliana by low phosphorus availability. Plant Cell Environ. *19*, 529–538.

Batista, P.J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., Bouley, D.M., Lujan, E., Haddad, B., Daneshvar, K., et al. (2014). m6A RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells. Cell Stem Cell *15*, 707–719.

Beckmann, B.M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., Schwarzl, T., Curk, T., Foehr, S., Huber, W., et al. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat. Commun. *6*, 10127.

Berkowitz, N.D., Silverman, I.M., Childress, D.M., Kazan, H., Wang, L.-S., and Gregory, B.D. (2016). A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer). BMC Bioinformatics *17*, 215.

Bernhardt, C., Lee, M.M., Gonzalez, A., Zhang, F., Lloyd, A., and Schiefelbein, J. (2003). The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root. Development *130*.

Bhaskaran, H., Rodriguez-Hernandez, A., and Perona, J.J. (2012). Kinetics of tRNA folding monitored by aminoacylation. RNA *18*, 569–580.

Blad, H., Reiter, N.J., Abildgaard, F., Markley, J.L., and Butcher, S.E. (2005). Dynamics and Metal Ion Binding in the U6 RNA Intramolecular Stem–Loop as Analyzed by NMR. J. Mol. Biol. *353*, 540–555.

Bodi, Z., Zhong, S., Mehra, S., Song, J., Graham, N., Li, H., May, S., and Fray, R.G. (2012). Adenosine Methylation in Arabidopsis mRNA is Associated with the 3′ End and Reduced Levels Cause Developmental Defects. Front. Plant Sci. *3*.

Bothe, J.R., Nikolova, E.N., Eichhorn, C.D., Chugh, J., Hansen, A.L., and Al-Hashimi, H.M. (2011). Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. Nat. Methods *8*, 919–931.

Brownlee, G.G., and Cartwright, E.M. (1977). Rapid gel sequencing of RNA by primed synthesis with reverse transcriptase. J. Mol. Biol. *114*, 93–117.

Brzezicha, B., Schmidt, M., Makalowska, I., Jarmolowski, A., Pienkowska, J., and Szweykowska-Kulinska, Z. (2006). Identification of human tRNA:m5C methyltransferase catalysing intron-dependent m5C formation in the first position of the anticodon of the pre-tRNA Leu (CAA). Nucleic Acids Res. *34*, 6034–6043.

Bullock, S.L., Ringel, I., Ish-Horowicz, D., and Lukavsky, P.J. (2010). A'-form RNA helices are required for cytoplasmic mRNA transport in Drosophila. Nat. Struct. Mol. Biol. *17*, 703–709.

Buratti, E., and Baralle, F.E. (2004). Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Mol. Cell. Biol. *24*, 10505–10514.

Cantara, W.A., Crain, P.F., Rozenski, J., McCloskey, J.A., Harris, K.A., Zhang, X., Vendeix, F.A.P., Fabris, D., and Agris, P.F. (2011). The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res. *39*, D195–D201.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., et al. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat. Genet. *43*, 956–963.

Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., and Gilbert, W.V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature *6*, 134–136.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642–655.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA Biology from an Atlas of Mammalian mRNA binding proteins. Cell *149*, 1393–1406.

Cech, T.R., and Steitz, J.A. (2014). The Noncoding RNA Revolution— Trashing Old Rules to Forge New Ones. Cell *157*, 77–94.

Chang, S.H., and RajBhandary, U.L. (1968). Studies on polynucleotides. LXXXI. Yeast phenylalanine transfer ribonucleic acid: partial digestion with pancreatic ribonuclease. J. Biol. Chem. *243*, 592–597.

Chapman, E.J., and Carrington, J.C. (2007). Specialization and evolution of endogenous small RNA pathways. Nat. Rev. Genet. *8*, 884–896.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol *10*, 741–754.

Chen, L., Bush, S.J., Tovar-Corona, J.M., Castillo-Morales, A., and Urrutia, A.O. (2014). Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol. Biol. Evol. *31*, 1402–1413.

Chen, T., Hao, Y.-J., Zhang, Y., Li, M.-M., Wang, M., Han, W., Wu, Y., Lv, Y., Hao, J., Wang, L., et al. (2015). m6A RNA Methylation Is Regulated by MicroRNAs and Promotes Reprogramming to Pluripotency. Cell Stem Cell *16*, 289–301.

Chiou, N.T., Shankarling, G., and Lynch, K.W. (2013). HnRNP L and HnRNP A1 Induce Extended U1 snRNA Interactions with an Exon to Repress Spliceosome Assembly. Mol Cell *49*, 972–982.

Clarke, J.H., Tack, D., Findlay, K., Van Montagu, M., and Van Lijsebettens, M. (1999). The SERRATE locus controls the formation of the early juvenile leaves and phase length in Arabidopsis. Plant J. Cell Mol. Biol. *20*, 493–501.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Cruz, J.A., and Westhof, E. (2009). The Dynamic Landscapes of RNA Architecture. Cell *136*, 604–609.

Darnell, J.E., Philipson, L., Wall, R., and Adesnik, M. (1971). Polyadenylic acid sequences: role in conversion of nuclear RNA into messenger RNA. Science *174*, 507–510.

Davis, F.F., and Allen, F.W. (1957). Ribonucleic acids from yeast which contain a fifth nucleotide. J. Biol. Chem. *227*, 907–915.

Deal, R.B., and Henikoff, S. (2010). A Simple Method for Gene Expression and Chromatin Profiling of Individual Cell Types within a Tissue. Dev. Cell *18*, 1030–1040.

Decatur, W.A., and Schnare, M.N. (2008). Different mechanisms for pseudouridine formation in yeast 5S and 5.8S rRNAs. Mol. Cell. Biol. *28*, 3089–3100.

Delatte, B., Wang, F., Ngoc, L.V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S., et al. (2016). RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. Science *351*, 282–285.

Demeshkina, N., Jenner, L., Yusupova, G., and Yusupov, M. (2010). Interactions of the ribosome with mRNA and tRNA. Curr. Opin. Struct. Biol. *20*, 325–332.

Desai, N.A., and Shankar, V. (2003). Single-strand-specific nucleases. FEMS Microbiol. Rev. *26*, 457–491.

Desrosiers, R., Friderici, K., and Rottman, F. (1974). Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. Proc. Natl. Acad. Sci. U. S. A. *71*, 3971–3975.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature *505*, 696–700.

Dolan, L., Janmaat, K., Willemsen, V., Linstead, P., Poethig, S., Roberts, K., and Scheres, B. (1993). Cellular organisation of the Arabidopsis thaliana root. Development *119*, 71–84.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201–206.

Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., and Rechavi, G. (2013). Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing. Nat. Protoc. *8*, 176–189.

Dominissini, D., Nachtergaele, S., Moshitch-Moshkovitz, S., Peer, E., Kol, N., Ben-Haim, M.S., Dai, Q., Di Segni, A., Salmon-Divon, M., Clark, W.C., et al. (2016). The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. Nature *530*, 441–446.

Dong, H., Ray, D., Ren, S., Zhang, B., Puig-Basagoiti, F., Takagi, Y., Ho, C.K., Li, H., and Shi, P.-Y. (2007). Distinct RNA Elements Confer Specificity to Flavivirus RNA Cap Methylation Events. J. Virol. *81*, 4412–4421.

Dong, Z., Han, M.-H., and Fedoroff, N. (2008). The RNA binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1. Proc. Natl. Acad. Sci. U. S. A. *105*, 9970–9975.

Draper, D.E. (2004). A guide to ions and RNA structure. RNA N. Y. N *10*, 335–343.

Draper, D.E. (2008). RNA Folding: Thermodynamic and Molecular Descriptions of the Roles of Ions. Biophys. J. *95*, 5489–5495.

Dreyfuss, G. (1986). Structure and function of nuclear and cytoplasmic ribonucleoprotein particles. Annu. Rev. Cell Biol. *2*, 459–498.

Dreyfuss, G., Choi, Y.D., and Adam, S.A. (1984). Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. Mol. Cell. Biol. *4*, 1104–1114.

Dreyfuss, G., Matunis, M.J., Piñol-Roma, S., and Burd, C.G. (1993). hnRNP proteins and the biogenesis of mRNA. Annu. Rev. Biochem. *62*, 289–321.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Z., S. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acid Res. *38*, W64–W70.

Ebhardt, H.A., Tsang, H.H., Dai, D.C., Liu, Y., Bostan, B., and Fahlman, R.P. (2009). Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. Nucleic Acids Res. *37*, 2461–2470.

Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.-P., and Ehresmann, B. (1987). Probing the structure of RNAs in solution. Nucleic Acid Res. *15*, 9109–9128.

Favorova, O.O., Fasiolo, F., Keith, G., Vassilenko, S.K., and Ebel, J.P. (1981). Partial digestion of tRNA--aminoacyl-tRNA synthetase complexes with cobra venom ribonuclease. Biochemistry (Mosc.) *20*, 1006–1011.

Fedoroff, N.V. (2002). RNA binding proteins in plants: the tip of an iceberg? Curr. Opin. Plant Biol. *5*, 452–459.

Fica, S.M., Tuttle, N., Novak, T., Li, N.-S., Lu, J., Koodathingal, P., Dai, Q., Staley, J.P., and Piccirilli, J.A. (2013). RNA catalyses nuclear pre-mRNA splicing. Nature *503*, 229–234.

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acid Res. *35*, W297–W299.

Foley, S.W., Vandivier, L.E., Kuksa, P.P., and Gregory, B.D. (2015). Transcriptome-wide measurement of plant RNA secondary structure. Curr. Opin. Plant Biol. *27*, 36–43.

Freeberg, M.A., Han, T., Moresco, J.J., Kong, A., Yang, Y.-C., Lu, Z.J., Yates, J.R., and Kim, J.K. (2013). Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae. Genome Biol. *14*, R13.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA binding proteins. Nat. Rev. Genet. *15*, 689–701.

Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m^6A RNA methylation. Nat. Rev. Genet. *15*, 293–306.

Furuta, K., Kubo, M., Sano, K., Demura, T., Fukuda, H., Liu, Y.-G., Shibata, D., and Kakimoto, T. (2011). The CKH2/PKL Chromatin Remodeling Factor Negatively Regulates Cytokinin Responses in Arabidopsis Calli. Pland Cell Physiol. *52*, 618–628.

Fustin, J.-M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., Isagawa, T., Morioka, M.S., Kakeya, H., Manabe, I., et al. (2013). RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell *155*, 793–806.

Ganot, P., Bortolin, M.-L., and Kiss, T. (1997). Site-Specific Pseudouridine Formation in Preribosomal RNA Is Guided by Small Nucleolar RNAs. Cell *89*, 799–809.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. Nat. Rev. Mol. Cell Biol. *8*, 113–126.

Gaston, K.W., and Limbach, P.A. (2014). The identification and characterization of non-coding and coding RNAs and their modified nucleosides by mass spectrometry. RNA Biol. *11*, 1568–1585.

Giegé, R., Jühling, F., Pütz, J., Stadler, P., Sauter, C., and Florentz, C. (2012). Structure of transfer RNAs: similarity and variability. Wiley Interdiscip. Rev. RNA *3*, 37–61.

Gilbert, G.A., Knight, J.D., Vance, C.P., and Allan, D.L. (1999). Acid phosphatase activity in phosphorus-deficient white lupin roots. Plant Cell Environ. *22*, 801–810.

Glisovic, T., Bachorik, J.L., Yong, J., and Dreyfuss, G. (2008). RNA binding proteins and post-transcriptional gene regulation. FEBS Lett. *582*, 1977–1986.

Goodarzi, H., Najafabadi, H.S., Oikonomou, P., Greco, T.M., Fish, L., Salavati, R., Cristea, I.M., and Tavazoie, S. (2012). Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature *485*, 264–268.

Gosai, S.J., Foley, S.W., Wang, D., Silverman, I.M., Selamoglu, N., Nelson, A.D.L., Beilstein, M.A., Daldal, F., Deal, R.B., and Gregory, B.D. (2015). Global Analysis of the RNA-Protein Interaction and RNA Secondary Structure Landscapes of the Arabidopsis Nucleus. Mol Cell *57*, 376–388.

Gregory, B.D., O'Malley, R.C., Lister, R., Urich, M.A., Tonti-Filippini, J., Chen, H., Millar, A.H., and Ecker, J.R. (2008). A link between RNA metabolism and silencing affecting Arabidopsis development. Dev. Cell *14*, 854–866.

Grierson, C., Nielsen, E., Ketelaarc, T., and Schiefelbein, J. (2014). Root Hairs. In The Arabidopsis Book, (The American Society of Plant Biologists), p. e0172.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. Nucleic Acids Res. *31*, 439–441.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., and Hofacker, I.L. (2008). The Vienna RNA Websuite. Nucleic Acids Res. *36*, W70–W74.

Grüter, P., Tabernero, C., von Kobbe, C., Schmitt, C., Saavedra, C., Bachi, A., Wilm, M., Felber, B.K., and Izaurralde, E. (1998). TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. Mol. Cell *1*, 649–659.

Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat. Struct. Mol. Biol. *21*, 198–206.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M.. J., Jungkamp, A.C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Han, H., Irimia, M., Ross, P.J., Sung, H.K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I.P., Nachman, E.N., et al. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. Nature *498*, 241–245.

Harris, K.A., Crothers, D.M., and Ullu, E. (1995). In vivo structural analysis of spliced leader RNAs in Trypanosoma brucei and Leptomonas collosoma: a flexible structure that is independent of cap4 methylations. RNA N. Y. N *1*, 351–362.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell *38*, 576–589.

Herlihy, A.E., and de Bruin, R.A.M. (2017). The Role of the Transcriptional Response to DNA Replication Stress. Genes *8*.

Holbrook, S.R., and Kim, S.H. (1997). RNA crystallography. Biopolymers *44*, 3–21.

Hoogstraten, C.G., Legault, P., and Pardi, A. (1998). NMR solution structure of the lead-dependent ribozyme: evidence for dynamics in RNA catalysis1. J. Mol. Biol. *284*, 337–350.

Hunt, A.G., Xing, D., and Li, Q.Q. (2012). Plant polyadenylation factors: conservation and variety in the polyadenylation complex in plants. BMC Genomics *13*, 641.

Hussain, S., Tuorto, F., Menon, S., Blanco, S., Cox, C., Flores, J.V., Watt, S., Kudo, N.R., Lyko, F., and Frye, M. (2013). The mouse cytosine-5 RNA methyltransferase NSun2 is a component of the chromatoid body and required for testis differentiation. Mol. Cell. Biol. *33*, 1561–1570.

Inoue, T., and Cech, T.R. (1985). Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: A technique for RNA structure analysis using chemical probes and reverse transcriptase. Proc Natl Acad Sci *82*, 648–652.

Izaurralde, E., Jarmolowski, A., Beisel, C., Mattaj, I.W., Dreyfuss, G., and Fischer, U. (1997). A role for the M9 transport signal of hnRNP A1 in mRNA nuclear export. J. Cell Biol. *137*, 27–35.

Jaffrey, S.R., and Kharas, M.G. (2017). Emerging links between m(6)A and misregulated mRNA methylation in cancer. Genome Med. *9*, 2.

Jangi, M., and Sharp, P.A. (2014). Building robust transcriptomes with master splicing factors. Cell *159*, 487–498.

Jelinek, W., Adesnik, M., Salditt, M., Sheiness, D., Wall, R., Molloy, G., Philipson, L., and Darnell, J.E. (1973). Further evidence on the nuclear origin and transfer to the cytoplasm of polyadenylic acid sequences in mammalian cell RNA. J. Mol. Biol. *75*, 515–532.

Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y.-G., et al. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat. Chem. Biol. *7*, 885–887.

Jin, Y., Yang, Y., and Zhang, P. (2011). New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. RNA Biol. *8*, 450–457.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K.V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim. Biophys. Acta *1840*, 1063–1071.

Kandasamy, M.K., McKinney, E.C., and Meagher, R.B. (1999). The late pollen-specific actins in angiosperms. Plant J. *18*, 681–691.

Kang, H., Park, S.J., and Kwak, K.J. (2013). Plant RNA chaperones in stress response. Trends Plant Sci. *18*, 100–106.

Kasai, H., Oashi, Z., Harada, F., Nishimura, S., Oppenheimer, N.J., Crain, P.F., Liehr, J.G., von Minden, D.L., and McCloskey, J.A. (1975). Structure of the modified nucleoside Q isolated from Escherichia coli transfer ribonucleic acid. 7-(4,5-cis-Dihydroxy-1-cyclopenten-3-ylaminomethyl)-7-deazaguanosine. Biochemistry (Mosc.) *14*, 4198–4208.

Keene, J.D., and Tenenbaum, S.A. (2002). Eukaryotic mRNPs May Represent Posttranscriptional Operons. Mol Cell *9*, 1151–1167.

Keene, J.D., Komisarow, J.M., and Friedersdorf, M.B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat. Protoc. *1*, 302–307.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

Kiledjian, M., DeMaria, C.T., Brewer, G., and Novick, K. (1997). Identification of AUF1 (heterogeneous nuclear ribonucleoprotein D) as a component of the alpha-globin mRNA stability complex. Mol. Cell. Biol. *17*, 4870–4876.

Kim, S.H., and Rich, A. (1968). Single crystals of transfer RNA: an x-ray diffraction study. Science *162*, 1381–1384.

Kim, J.Y., Park, S.J., Jang, B., Jung, C.-H., Ahn, S.J., Goh, C.-H., Cho, K., Han, O., and Kang, H. (2007). Functional characterization of a glycine-rich RNA binding protein 2 in Arabidopsis thaliana under abiotic stress conditions. Plant J. Cell Mol. Biol. *50*, 439–451.

Kim, S.H., Suddath, F.L., Quigley, G.J., McPherson, A., Sussman, J.L., Wang, A.H., Seeman, N.C., and Rich, A. (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science *185*, 435–440.

Kim, Y.O., Kim, J.S., and Kang, H. (2005). Cold-inducible zinc finger-containing glycine-rich RNA binding protein contributes to the enhancement of freezing tolerance in Arabidopsis thaliana. Plant J. *42*, 890–900.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA binding proteins. Nat. Methods *8*, 559–564.

Kiss, T. (2002). Small Nucleolar RNAs: An Abundant Group of Noncoding RNAs with Diverse Cellular Functions. Cell *109*, 145–148.

Kiss-László, Z., Henry, Y., Bachellerie, J.-P., Caizergues-Ferrer, M., and Kiss, T. (1996). Site-Specific Ribose Methylation of Preribosomal RNA: A Novel Function for Small Nucleolar RNAs. Cell *85*, 1077–1088.

Klasens, B.I., Das, A.T., and Berkhout, B. (1998). Inhibition of polyadenylation by stable RNA secondary structure. Nucleic Acids Res. *26*, 1870–1876.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Knapp, G. (1989). Enzymatic approaches to probing of RNA secondary and tertiary structure. Methods Enzymol. *180*, 192–212.

Köhler, A., and Hurt, E. (2007). Exporting RNA from the nucleus to the cytoplasm. Nat. Rev. Mol. Cell Biol. *8*, 761–773.

Kortmann, J., and Narberhaus, F. (2012). Bacterial RNA thermometers: molecular zippers and switches. Nat. Rev. Microbiol. *10*, 255–265.

Kozak, M. (1988). Leader length and secondary structure modulate mRNA function under conditions of stress. Mol. Cell. Biol. *8*, 2737–2744.

Kuninaka, A., Kibi, M., Yoshino, H., and Sakaguchi, K. (1961). Studies on 5′-Phosphodiesterases in Microorganisms. Agric. Biol. Chem. *25*, 693–701.

Kupsch, C., Ruwe, H., Gusewski, S., Tillich, .M., Small, I., and Schmitz-Linneweber, C. (2012). Arabidopsis Chloroplast RNA Binding Proteins CP31A and CP29A Associate with Large Transcript Pools and Confer Cold Stress Tolerance by Influencing Multiple Chloroplast RNA Processing Steps. Plant Cell *24*, 4266–4280.

Kurihara, Y., and Watanabe, Y. (2004). Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. Proc. Natl. Acad. Sci. U. S. A. *101*, 12753–12758.

Kwak, K.J., Kim, Y.O., and Kang, H. (2005). Characterization of transgenic Arabidopsis plants overexpressing GR-RBP4 under high salinity, dehydration, or cold stress. J Exp Bot *56*, 3007–3016.

Lambert, D., and Draper, D.E. (2007). Effects of Osmolytes on RNA Secondary and Tertiary Structure Stabilities and RNA-Mg2+ Interactions. J. Mol. Biol. *370*, 993–1005.

Lan, P., Li, W., Lin, W.-D., Santi, S., and Schmidt, W. (2013). Mapping gene activity of Arabidopsis root hairs. Genome Biol. *14*, R67.

Laubinger, S., Sachsenberg, T., Zeller, G., Busch, W., Lohmann, J.U., Rätsch, G., and Weigel, D. (2008). Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. *105*, 8795–8800.

Lawley, P.D., and Brookes, P. (1963). FURTHER STUDIES ON THE ALKYLATION OF NUCLEIC ACIDS AND THEIR CONSTITUENT NUCLEOTIDES. Biochem. J. *89*, 127–138.

Lebedeva, S., Jens, M., Theil, K., Schwanhausser, B., Selbach, M., Landthaler, M., and Rajewsky, N. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA binding protein HuR. Mol Cell *43*, 340–352.

Lee, M.M., and Schiefelbein, J. (2002). Cell pattern in the Arabidopsis root epidermis determined by lateral inhibition with feedback. Plant Cell *14*, 611–618.

Lee, M.S., Henry, M., and Silver, P.A. (1996). A protein that shuttles between the nucleus and the cytoplasm is an important mediator of RNA export. Genes Dev. *10*, 1233–1246.

Lempereur, L., Nicoloso, M., Riehl, N., Ehresmann, C., Ehresmann, B., and Bachellerie, J.P. (1985). Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible. Nucleic Acids Res. *13*, 8339–8357.

Leontis, N.B., and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. RNA *7*, 499–512.

Lestrade, L., and Weber, M.J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. *34*, D158–D162.

Lewinski, M., Hallmann, A., and Staiger, D. (2016). Genome-wide identification and phylogenetic analysis of plant RNA binding proteins comprising both RNA recognition motifs and contiguous glycine residues. Mol. Genet. Genomics MGG *291*, 763–773.

Li, F., Zheng, Q., Ryvkin, P., Dragomir, I., Desai, Y., Aiyer, S., Valladares, O., Yang, J., Bambina, S., Sabin, L.R., et al. (2012a). Global Analysis of RNA Secondary Structure in Two Metazoans. Cell Rep. *1*, 69–82.

Li, F., Zheng, Q., Vandivier, L.E., Willmann, M.R., Chen, Y., and Gregory, B.D. (2012b). Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome. Plant Cell *24*, 4346–4359.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Curr. Biol. CB *15*, 1501–1507.

Li, S., Yamada, M., Han, X., Ohler, U., and Benfey, P.N. (2016). High-Resolution Expression Map of the Arabidopsis Root Reveals Alternative Splicing and lincRNA Regulation. Dev. Cell *39*, 508–522.

Li, X., Zhu, P., Ma, S., Song, J., Bai, J., Sun, F., and Yi, C. (2015). Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. Nat. Chem. Biol. *11*, 592–597.

Licatalosi, D.D., A., M., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Limbach, P.A., Crain, P.F., and McCloskey, J.A. (1994). Summary: the modified nucleosides of RNA. Nucleic Acids Res. *22*, 2183–2196.

Linder, B., Grozhik, A.V., Olarerin-George, A.O., Meydan, C., Mason, C.E., and Jaffrey, S.R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat. Methods *12*, 767–772.

Linkohr, B.I., Williamson, L.C., Fitter, A.H., and Leyser, H.M.O. (2002). Nitrate and phosphate availability and distribution have different effects on root system architecture of Arabidopsis. Plant J. *29*, 751–760.

Liu, H.X., Goodall, G.J., Kole, R., and Filipowicz, W. (1995). Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in Nicotiana plumbaginifolia. EMBO J. *14*, 377–388.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.H. (2012). Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis. Plant Cell *24*, 4333–4345.

Liu, J., Yue, Y., Han, D., Wang, X., Fu, Y., Zhang, L., Jia, G., Yu, M., Lu, Z., Deng, X., et al. (2014). A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat. Chem. Biol. *10*, 93–95.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. Nature *518*, 560–564.

Lockard, R.E., and Kumar, A. (1981). Mapping tRNA structure in solution using double-strand-specific ribonuclease V1 from cobra venom. Nucleic Acids Res. *9*, 5125–5140.

Lorković, Z.J. (2009). Role of plant RNA binding proteins in development, stress response and genome organization. Trends Plant Sci. *14*, 229–236.

Lorković, Z.J., and Barta, A. (2002). Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA binding proteins from the flowering plant Arabidopsis thaliana. Nucleic Acids Res. *30*, 623–635.

Lorsch, J.R. (2002). RNA Chaperones Exist and DEAD Box Proteins Get a Life. Cell *109*, 797–800.

Lovejoy, A.F., Riordan, D.P., and Brown, P.O. (2014). Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. PloS One *9*, e110799.

Loverix, S., and Steyaert, J. (2001). Deciphering the mechanism of RNase T1. Methods Enzymol. *341*, 305–323.

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. *53*, 661–673.

Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K.M., et al. (2013). MODOMICS: a database of RNA modification pathways--2013 update. Nucleic Acids Res. *41*, D262–D267.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Madhani, H.D. (2013). snRNA Catalysts in the Spliceosome's Ancient Core. Cell *155*, 1213–1215.

Marondedze, C., Thomas, L., Serrano, N.L., Lilley, K.S., and Gehring, C. (2016). The RNA binding protein repertoire of Arabidopsis thaliana. Sci. Rep. *6*, 29766.

Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. *22*, 1184–1195.

Mathews, D.H. (2014). RNA Secondary Structure Analysis Using RNAstructure. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *46*, 12.6.1–12.6.25.

Matia-González, A.M., Laing, E.E., and Gerber, A.P. (2015). Conserved mRNA-binding proteomes in eukaryotic organisms. Nat. Struct. Mol. Biol. *22*, 1027–1033.

Mayrand, S., Setyono, B., Greenberg, J., and Pederson, T. (1981). Structure of nuclear ribonucleoprotein: identification of proteins in contact with poly(A)+ heterogeneous nuclear RNA in living HeLa cells. J. Cell Biol. *90*, 380–384.

McCloskey, J.A., and Rozenski, J. (2005). The Small Subunit rRNA Modification Database. Nucleic Acids Res. *33*, D135–D138.

McHugh, C.A., Russell, P., and Guttman, M. (2014). Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. *15*, 203.

McKee, A.E., Minet, E., Stern, C., Riahi, S., Stiles, C.D., and Silver, P.A. (2005). A genome-wide in situ hybridization map of RNA binding proteins reveals anatomically restricted expression in the developing mouse brain. BMC Dev. Biol. *5*, 14.

Mei, H., Cheng, N.H., Zhao, J., Park, S., Escareno, R.A., Pittman, J.K., and Hirschi, K.D. (2009). Root development under metal stress in Arabidopsis thaliana requires the H+/cation antiporter CAX4. New Phytol. *183*, 95–105.

Meisner, C.A., and Karnok, K.J. (1991). Root Hair Occurrence anld Variation with Environment. Agron. J. *83*, 814.

Meng, Z., and Limbach, P.A. (2006). Mass spectrometry of RNA: linking the genome to the proteome. Brief. Funct. Genomic. Proteomic. *5*, 87–95.

Merino, E.J., Wilkinson, K.A., Coughlan, J.L., and Weeks, K.M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J. Am. Chem. Soc. *127*, 4223–4231.

Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell *149*, 1635–1646.

Mohr, S., Stryker, J.M., and Lambowitz, A.M. (2002). A DEAD-Box Protein Functions as an ATP-Dependent RNA Chaperone in Group I Intron Splicing. Cell *109*, 769–779.

Motorin, Y., Muller, S., Behm-Ansmant, I., and Branlant, C. (2007). Identification of modified residues in RNAs by reverse transcription-based methods. Methods Enzymol. *425*, 21–53.

Muchhal, U.S., Pardo, J.M., and Raghothama, K.G. (1996). Phosphate transporters from the higher plant Arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. *93*, 10519–10523.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Nakazato, H., Kopp, D.W., and Edmonds, M. (1973). Localization of the polyadenylate sequences in messenger ribonucleic acid and in the heterogeneous nuclear ribonucleic acid of HeLa cells. J. Biol. Chem. *248*, 1472–1476.

Newby, M.I., and Greenbaum, N.L. (2002). Investigation of Overhauser effects between pseudouridine and water protons in RNA helices. Proc. Natl. Acad. Sci. U. S. A. *99*, 12697–12702.

Ni, J., Tien, A.L., and Fournier, M.J. (1997). Small Nucleolar RNAs Direct Site-Specific Synthesis of Pseudouridine in Ribosomal RNA. Cell *89*, 565–573.

Nissen, P., Hansen, J., Ban, N., Moore, P.B., and Steitz, T.A. (2000). The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. Science *289*, 920–930.

Niu, Y.F., Chai, R.S., Jin, G.L., Wang, H., Tang, C.X., and Zhang, Y.S. (2013). Responses of root architecture development to low phosphorus availability: a review. Ann. Bot. *112*, 391–408.

Novikova, I.V., Hennelly, S.P., and Sanbonmatsu, K.Y. (2012). Sizing up long non-coding RNAs: Do lncRNAs have secondary and tertiary structure? BioArchitecture *2*, 189–199.

OECD, and FAO (2012). OECD-FAO Agricultural Outlook 2012 (Paris: Organisation for Economic Co-operation and Development).

Ofengand, J., and Bakin, A. (1997). Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. J. Mol. Biol. *266*, 246–268.

Ogas, J., Kaufmann, S., Henderson, J., and Somerville, C. (1999). PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis. PNAS *96*, 13839–13844.

Oikawa, D., Tokuda, M., Hosoda, A., and Iwawaki, T. (2010). Identification of a consensus element recognized and cleaved by IRE1 alpha. Nucleic Acids Res. *38*, 6265–6273.

Park, W., Li, J., Song, R., Messing, J., and Chen, X. (2002). CARPEL FACTORY, a Dicer Homolog, and HEN1, a Novel Protein, Act in microRNA Metabolism in Arabidopsis thaliana. Curr. Biol. *12*, 1484–1495.

Patrick, W.H., and Khalid, R.A. (1974). Phosphate Release and Sorption by Soils and Sediments: Effect of Aerobic and Anaerobic Conditions. Science *186*, 53–55.

Peattie, D.A. (1979). Direct chemical method for sequencing RNA. Proc Natl Acad Sci *76*, 1760–1764.

Peattie, D.A., and Gilbert, W. (1980). Chemical probes for higher-order structure in RNA. Proc Natl Acad Sci *77*, 4679–4682.

Péret, B., Clément, M., Nussaume, L., and Desnos, T. (2011). Root developmental adaptation to phosphate starvation: better safe than sorry. Trends Plant Sci. *16*, 442–450.

Piñol-Roma, S., Choi, Y.D., Matunis, M.J., and Dreyfuss, G. (1988). Immunopurification of heterogeneous nuclear ribonucleoprotein particles reveals an assortment of RNA binding proteins. Genes Dev. *2*, 215–227.

Ponting, C.P., Oliver, P.L., and Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. Cell *136*, 629–641.

Raczynska, K.D., Stepien, A., Kierzkowski, D., Kalak, M., Bajczyk, M., McNicol, J., Simpson, C.G., Szweykowska-Kulinska, Z., Brown, J.W.S., and Jarmolowski, A. (2014). The SERRATE protein is involved in alternative splicing in Arabidopsis thaliana. Nucleic Acids Res. *42*, 1224–1244.

Rajagopalan, L.E., Westmark, C.J., Jarzembowski, J.A., and Malter, J.S. (1998). hnRNP C increases amyloid precursor protein (APP) production by stabilizing APP mRNA. Nucleic Acids Res. *26*, 3418–3423.

Raker, V.A., Mironov, A.A., Gelfand, M.S., and Pervouchine, D.D. (2009). Modulation of alternative splicing by long-range RNA structures in Drosophila. Nucleic Acid Res. *37*, 4533–4544.

Ramakrishnan, V. (2014). The Ribosome Emerges from a Black Box. Cell *159*, 979–984.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. (2002). MicroRNAs in plants. Genes Dev. *16*, 1616–1626.

Reynolds, N., and Cooke, H.J. (2005). Role of the DAZ genes in male fertility. Reprod. Biomed. Online *10*, 72–80.

Robertus, J.D., Ladner, J.E., Finch, J.T., Rhodes, D., Brown, R.S., Clark, B.F., and Klug, A. (1974). Structure of yeast phenylalanine tRNA at 3 A resolution. Nature *250*, 546–551.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Rux, D.R., and Wellik, D.M. (2016). Hox genes in the adult skeleton: Novel functions beyond embryonic development. Dev. Dyn. Off. Publ. Am. Assoc. Anat.

Ryu, K.H., Kang, Y.H., Park, Y., Hwang, I., Schiefelbein, J., and Lee, M.M. (2005). The WEREWOLF MYB protein directly regulates CAPRICE transcription during cell fate specification in the Arabidopsis root epidermis. Dev. Camb. Engl. *132*, 4765–4775.

Ryvkin, P., Leung, Y.Y., Silverman, I.M., Childress, M., Valladares, O., Dragomir, I., Gregory, B.D., and Wang, L.S. (2013). HAMR: high-throughput annotation of modified ribonucleotides. RNA *19*, 1684–1692.

Saletore, Y., Meyer, K., Korlach, J., Vilfan, I.D., Jaffrey, S., and Mason, C.E. (2012). The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol. *13*, 175.

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., Spence, M., et al. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. eLife *2*, e01749.

Scherrer, T., Mittal, N., Janga, S.C., and Gerber, A.P. (2010). A screen for RNA binding proteins in yeast indicates dual functions for many enzymes. PloS One *5*, e15499.

Schiefelbein, J. (2003). Cell-fate specification in the epidermis: a common patterning mechanism in the root and shoot. Curr. Opin. Plant Biol. *6*, 74–78.

Schmitz-Linneweber, C., Williams-Carrier, R., and Barkan, A. (2005). RNA immunoprecipitation and microarray analysis show a chloroplast Pentatricopeptide repeat protein to be associated with the 5' region of mRNAs whose translation it activates. Plant Cell *17*, 2791–2804.

Schmucker, D., Clemens, J.C., Shu, H., Worby, C.A., Xiao, J., Muda, M., Dixon, J.E., and Zipursky, S.L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell *101*, 671–684.

Schöning, J.C., Streitner, C., Meyer, I.M., Gao, Y., and Staiger, D. (2008). Reciprocal regulation of glycine-rich RNA binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. Nucleic Acids Res. *36*, 6977–6987.

Schroeder, R., Barta, A., and Semrad, K. (2004). Strategies for RNA folding and assembly. Nat. Rev. Mol. Cell Biol. *5*, 908–919.

Schwartz, S., Mumbach, M.R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G.G., Mertins, P., Ter-Ovanesyan, D., Habib, N., Cacchiarelli, D., et al. (2014a). Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites. Cell Rep. *8*, 284–296.

Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., León-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S., et al. (2014b). Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA. Cell.

Scott, M.P. (2016). Homeodomains, Hedgehogs, and Happiness. Curr. Top. Dev. Biol. *117*, 331–337.

Serrano-Cartagena, J., Robles, P., Ponce, M.R., and Micol, J.L. (1999). Genetic analysis of leaf form mutants from the Arabidopsis Information Service collection. Mol. Gen. Genet. MGG *261*, 725–739.

Shang, X., Cao, Y., and Ma, L. (2017). Alternative Splicing in Plant Genes: A Means of Regulating the Environmental Fitness of Plants. Int. J. Mol. Sci. *18*.

Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Ozsolak, F., Milos, P.M., Barton, G.J., and Simpson, G.G. (2012). Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nat. Struct. Mol. Biol. *19*, 845–852.

Silverman, I.M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J.L., and Gregory, B.D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. *15*, R3.

Simpson, C.G., Manthri, S., Raczynska, K.D., Kalyna, M., Lewandowska, D., Kusenda, B., Maronova, M., Szweykowska-Kulinska, Z., Jarmolowski, A., Barta, A., et al. (2010). Regulation of plant gene expression by alternative splicing. Biochem Soc Trans. *38*, 667–671.

Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., et al. (2015). Structural imprints in vivo decode RNA regulatory mechanisms. Nature *519*, 486–490.

Sprinzl, M., and Vassilenko, K.S. (2005). Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. *33*, D139–D140.

Steitz, T.A., and Moore, P.B. (2003). RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem. Sci. *28*, 411–418.

Streitner, C., Köster, T., Simpson, C.G., Shaw, P., Danisman, S., Brown, J.W.S., and Staiger, D. (2012). An hnRNP-like RNA binding protein affects alternative splicing by in vivo interaction with transcripts in Arabidopsis thaliana. Nucleic Acids Res. *40*, 11240–11255.

Subramanian, M., Rage, F., Tabet, R., Flatter, E., Mandel, J.-L., and Moine, H. (2011). G-quadruplex RNA structure as a signal for neurite mRNA targeting. EMBO Rep. *12*, 697–704.

Svitkin, Y.V., Pause, A., Haghighat, A., Pyronnet, S., Witherell, G., Belsham, G.J., and Sonenberg, N. (2001). The requirement for eukaryotic initiation factor 4A (eIF4A) in translation is in direct proportion to the degree of mRNA 5' secondary structure. RNA N. Y. N *7*, 382–394.

Talkish, J., May, G., Lin, Y., Woolford Jr., J.L., and McManus, C.J. (2014). Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA *20*, 1–8.

Tenenbaum, S.A., Carson, C.C., Lager, P.J., and Keene, J.D. (2000). Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc. Natl. Acad. Sci. U. S. A. *97*, 14085–14090.

Tenenbaum, S.A., Christiansen, J., and Nielsen, H. (2011). The post-transcriptional operon. Methods Mol Biol *703*, 237–245.

Tompa, P., and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. FASEB J. *18*, 1169–1175.

Tsvetanova, N.G., Klass, D.M., Salzman, J., and Brown, P.O. (2010). Proteome-wide search reveals unexpected RNA binding proteins in Saccharomyces cerevisiae. PloS One *5*.

Uchida, T., Arima, T., and Egami, F. (1970). Specificity of RNase U2. J. Biochem. (Tokyo) *67*, 91–102.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. Science *302*, 1212–1215.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat. Methods *7*, 995–1001.

Vandivier, L., Li, F., Zheng, Q., Willmann, M., Chen, Y., and Gregory, B. (2013). Arabidopsis mRNA secondary structure correlates with protein function and domains. Plant Signal. Behav. *8*, e24301.

Vandivier, L.E., Campos, R., Kuksa, P.P., Silverman, I.M., Wang, L.-S., and Gregory, B.D. (2015). Chemical Modifications Mark Alternatively Spliced and Uncapped Messenger RNAs in Arabidopsis. Plant Cell *27*, 3024–3037.

Vandivier, L.E., Anderson, S.J., Foley, S.W., and Gregory, B.D. (2016). The Conservation and Function of RNA Secondary Structure in Plants. Annu. Rev. Plant Biol. *67*, 463–488.

van Eekelen, C.A., Riemen, T., and van Venrooij, W.J. (1981). Specificity in the interaction of hnRNA and mRNA with proteins as revealed by in vivo cross linking. FEBS Lett. *130*, 223–226.

Vishwakarma, K., Upadhyay, N., Kumar, N., Yadav, G., Singh, J., Mishra, R.K., Kumar, V., Verma, R., Upadhyay, R.G., Pandey, M., et al. (2017). Abscisic Acid Signaling and Abiotic Stress Tolerance in Plants: A Review on Current Knowledge and Future Prospects. Front. Plant Sci. *8*, 161.

Volkin, E., and Cohn, W.E. (1953). On the structure of ribonucleic acids. II. The products of ribonuclease action. J. Biol. Chem. *205*, 767–782.

Wada, T., Tachibana, T., Shimura, Y., and Okada, K. (1997). Epidermal cell differentiation in Arabidopsis determined by a Myb homolog, CPC. Science *277*, 1113–1116.

Wahl, M.C., Will, C.L., and Luhrmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell *136*, 701–718.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706–709.

Wan, Y., Tang, K., Zhang, D., Xie, S., Zhu, X., Wang, Z., and Lang, Z. (2015). Transcriptome-wide high-throughput deep m6A-seq reveals unique differential m6A methylation patterns between three organs in Arabidopsis thaliana. Genome Biol. *16*, 272.

Wang, D., and Deal, R.B. (2015). Epigenome profiling of specific plant cell types using a streamlined INTACT protocol and ChIP-seq. Methods Mol. Biol. Clifton NJ *1284*, 3–25.

Wang, K.C., and Chang, H.Y. (2011). Molecular Mechanisms of Long Noncoding RNAs. Mol. Cell *43*, 904–914.

Wang, H., Xu, Q., Kong, Y.-H., Chen, Y., Duan, J.-Y., Wu, W.-H., and Chen, Y.-F. (2014). Arabidopsis WRKY45 transcription factor activates PHOSPHATE TRANSPORTER1;1 expression in response to phosphate starvation. Plant Physiol. *164*, 2020–2029.

Wang, X., Zhao, B.S., Roundtree, I.A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. Cell *161*, 1388–1399.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. *35*, 169–178.

Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B., and Carstens, R.P. (2009). ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. Mol. Cell *33*, 591–601.

Wei, W., Ji, X., Guo, X., and Ji, S. (2017). Regulatory Role of N(6) -Methyladenosine (m(6) A) Methylation in RNA Processing and Human Diseases. J. Cell. Biochem.

Wells, S.E., Hughes, J.M., Igel, A.H., and Ares, M. (2000). Use of dimethyl sulfate to probe RNA structure in vivo. Methods Enzymol. *318*, 479–493.

Wen, J.-D., Lancaster, L., Hodges, C., Zeri, A.-C., Yoshimura, S.H., Noller, H.F., Bustamante, C., and Tinoco, I. (2008). Following translation by single ribosomes one codon at a time. Nature *452*, 598–603.

Wetzel, C., and Limbach, P.A. (2016). Mass spectrometry of modified RNAs: recent developments. The Analyst *141*, 16–23.

Whittington, A.T., Vugrek, O., Wei, K.J., Hasenbein, N.G., Sugimoto, K., Rashbrooke, M.C., and Wasteneys, G.O. (2001). MOR1 is essential for organizing cortical microtubules in plants. Nature *411*, 610–613.

Wilkinson, K.A., Merino, E.J., and Weeks, K.M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat. Protoc. *1*, 1610–1616.

Williamson, L.C., Ribrioux, S.P.C.P., Fitter, A.H., and Leyser, H.M.O. (2001). Phosphate Availability Regulates Root System Architecture in Arabidopsis. Plant Physiol. *126*, 875–882.

Willmann, M.R., Berkowitz, N.D., and Gregory, B.D. (2014). Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes—GMUCT 2.0. Methods *67*, 64–73.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. *37*, D380–D386.

Woo, J., MacPherson, C.R., Liu, J., Wang, H., Kiba, T., Hannah, M.A., Wang, X.-J., Bajic, V.B., and Chua, N.-H. (2012). The response and recovery of the Arabidopsis thalianatranscriptome to phosphate starvation. BMC Plant Biol. *12*, 62.

Wu, H.P., Su, Y.S., Chen, H.C., Chen, Y.R., Wu, C.C., Lin, W.D., and Tu, S.L. (2014). Genome-wide analysis of light-regulated alternative splicing mediated by photoreceptors in Physcomitrella patens. Genome Biol. *15*, R10.

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q.Q., and Huntd, A.G. (2011). Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. PNAS *108*, 12533–12538.

Xiao, W., Adhikari, S., Dahal, U., Chen, Y.-S., Hao, Y.-J., Sun, B.-F., Sun, H.-Y., Li, A., Ping, X.-L., Lai, W.-Y., et al. (2016). Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. Mol. Cell *61*, 507–519.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol. Cell *36*, 996–1006.

Yang, L., Liu, Z., Lu, F., Dong, A., and Huang, H. (2006). SERRATE is a novel nuclear regulator in primary microRNA processing in Arabidopsis. Plant J. Cell Mol. Biol. *47*, 841–850.

Ye, L., Li, Y., Fukami-Kobayashil, K., Go, M., Konishi, T., Watanabe, A., and Sugiura, M. (1991). Diversity of a ribonucleoprotein family in tobacco chloroplasts: two new chloroplast ribonucleoproteins and a phylogenetic tree of ten chloroplast RNA-binding domains. Nucleic Acid Res. *19*, 6485–6490.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.-D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat. Struct. Mol. Biol. *16*, 130–137.

Younis, I., Dittmar, K., Wang, W., Foley, S.W., Berg, M.G., Hu, K.Y., Wei, Z., Wan, L., and Dreyfuss, G. (2013). Minor introns are embedded molecular switches regulated by highly unstable U6atac snRNA. eLife *2*, e00780.

Yusupova, G., and Yusupov, M. (2014). High-Resolution Structure of the Eukaryotic 80S Ribosome. Annu. Rev. Biochem. *83*, 467–486.

Zaug, A.J., and Cech, T.R. (1995). Analysis of the structure of Tetrahymena nuclear RNAs in vivo: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. RNA N. Y. N *1*, 363–374.

Zhang, Y., Wang, X., Lu, S., and Liu, D. (2014). A major root-associated acid phosphatase in Arabidopsis, AtPAP10, is regulated by both local and systemic signals under phosphate starvation. J. Exp. Bot. *65*, 6577–6588.

Zhao, B., and Zhang, Q. (2015). Characterizing excited conformational states of RNA by NMR spectroscopy. Curr. Opin. Struct. Biol. *30*, 134–146.

Zhao, B.S., Roundtree, I.A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. Nat. Rev. Mol. Cell Biol. *18*, 31–42.

Zhao, X., Yang, Y., Sun, B.-F., Shi, Y., Yang, X., Xiao, W., Hao, Y.-J., Ping, X.-L., Chen, Y.-S., Wang, W.-J., et al. (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell Res. *24*, 1403–1419.

Zheng, D., and Tian, B. (2014). RNA binding proteins in regulation of alternative cleavage and polyadenylation. Adv. Exp. Med. Biol. *825*, 97–127.

Zheng, G., Dahl, J.A., Niu, Y., Fedorcsak, P., Huang, C.-M., Li, C.J., Vågbø, C.B., Shi, Y., Wang, W.-L., Song, S.-H., et al. (2013). ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol. Cell *49*, 18–29.

Zheng, Q., Ryvkin, P., Li, F., Dragomir, I., Valladares, O., Yang, J., Cao, K., Wang, L.S., and Gregory, B.D. (2010). Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in Arabidopsis. PLoS Genet. *6*, e1001141.

Zhou, H., Kimsey, I.J., Nikolova, E.N., Sathyamoorthy, B., Grazioli, G., McSally, J., Bai, T., Wunderlich, C.H., Kreutz, C., Andricioaei, I., et al. (2016). m1A and m1G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. Nat. Struct. Mol. Biol. *23*, 803–810.

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. *9*, 133–148.