IDENTIFYING RNA-PROTEIN INTERACTION SITES THROUGHOUT EUKARYOTIC TRANSCRIPTOMES

Ian Michael Silverman

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

2015

Supervisor of Dissertation

_____

Brian D. Gregory
Assistant Professor of Biology

Graduate Group Chairperson

_____

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

Dissertation Committee

Kristen W. Lynch, Ph.D. (Chair)
Professor of Biochemistry and Biophysics

Nancy M. Bonini, Ph.D.
Florence R.C. Murray Professor of Biology

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

Christopher D. Brown
Assistant Professor of Genetics

IDENTIFYING RNA-PROTEIN INTERACTION SITES THROUGHOUT EUKARYOTIC TRANSCRIPTOMES

COPYRIGHT

2015

Ian Michael Silverman

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License

To view a copy of this license, visit

http://creativecommons.org/licenses/by-nc-sa/2.0/

ACKNOWLEDGMENTS

The contents of this dissertation would not have been possible without the assistance and support of so many people. I am forever grateful to my advisor Brian Gregory, who has ferociously lead and mentored me over the past 5 years. Thank you for teaching me “to go where the biology takes me” and that a failed experiment is just the start of a new project. Thank you also to my committee members, who have guided me through the Ph.D. process and encouraged me to pursue multiple research interests during my training. I would also like to thank my undergraduate mentors; David Davies and Claudia Marques at Binghamton University and Jens Gundlach and Marcus Collins at the University of Washington. Thanks for giving me the opportunity to do exciting science and inspiring me to pursue a graduate degree.

There is no doubt that I am indebted to my labmates, both former and current; Isabelle Dragomir and Matthew Willmann, for teaching me to work with RNA and basic molecular biology techniques; Fan Li, Nate Berkowitz and Sager Gosai for being patient with me as I learned programming and bioinformatics skills and for being excellent collaborators and friends. Thanks also to Lee Vandivier, Qi Zheng, Shawn Foley, and Xiang Yu, Anissa Alexander, Steve Anderson and Lucy Shan for your support and for keeping the lab a fun and interesting place to work.

My dissertation research could not have been completed without the assistance of members of other laboratories. Thanks go to John Rinn, Cole Trapnell and Loyal Goff, who helped to develop the concept of PIP-seq and initiated the experiments. I’d also like to thank members of the Liebhaber lab, especially Hemant Kini, Xinjun Ji and Lou Ghanem for being excellent collaborators and also mentoring me during my time at Penn. I look forward to our future collaborations. Thanks also to members of the Mourelatos lab including Nikolaus Vrettos and Anastasios (Tassos) Vourekas for their assistance with microRNA immunoprecipitation experiments.

In addition to my dissertation work, I have had the honor of working with many excellent collaborators from a wide array of laboratories, both at Penn and other institutions. Thanks go to

Li-San Wang and Paul Ryvkin, for letting me contribute to their project on RNA modifications. Many thanks to Muredach Reilly, Mingyao Li and their lab members, especially Hanrui Zhang and Yichuan Liu for collaborative work on transcriptome profiling in human blood and adipose tissue. It was a pleasure working with David Fredrick and Joseph Baur on transcriptome profiling of a mouse model of muscular dystrophy. Thanks to Xiafeng Cao and Toalan Zhao for allowing me to analyze one of the first *Arabidopsis* CLIP-seq datasets. Also thanks to Henry Daniell, John Cupp and Elise Van Buskirk for working with us on transcriptome analysis of squalene producing Tobacco plants.

All of my work has been dependent on the hard work and dedication of two high-throughput sequencing cores. The late Penn Genome Frontiers Institute sequencing core, run by Jeanne Geskes and staffed by Davinder Sandu and Amber Kiliti. Also, thanks to the Next Generation Sequencing Core, their tireless leader Jonathan Schug, and members of his staff, especially Allan Fox, Joe Grubb, Christina Theodorou and Haleigh Zillges. Many thanks to the CAMB coordinators, especially Meagan Schofer, for keeping our graduate program running like a well-oiled machine.

I would also like to thank my classmates and friends in GGR, CAMB and other programs at Penn. You have made graduate school an amazing experience and I've learned so much from each of you. Thanks to all of my non-science friends who have supported my interest in science and pretended to be interested when I discussed RNA at length.

I must thank my family for their love and support. To my parents, Sue and Dan, thank you for encouraging me to ask questions and providing me with endless opportunities to learn. Thanks to my brother Eric, and his wife Lauren, for setting an example and demonstrating that hard work and dedication can pay off. Thanks also to my sister Julie, who showed me that an interest in science is not equivalent to an interest in medicine and for paving the way for my scientific career.

Katie, thank you for keeping me balanced through my time in graduate school. I don't know how I would have made it through alone, but I'm happy I don't have to find out.

ABSTRACT

IDENTIFYING RNA-PROTEIN INTERACTION SITES THROUGHOUT EUKARYOTIC TRANSCRIPTOMES

Ian Michael Silverman

Brian D. Gregory, Ph.D.

Gene expression is regulated at both the transcriptional and post-transcriptional levels. While transcription controls only the rate of RNA production, numerous and diverse mechanisms regulate the processing, stability and translation of RNAs at the post-transcriptional level. At the heart of this regulation are RNA-binding proteins (RBPs) and their RNA targets. Thousands of RBPs are encoded in mammalian genomes, each with hundreds to thousands of RNA targets. Therefore, cataloging these interactions represents a significant challenge. Recent advances in high-throughput sequencing technologies have greatly expanded the toolkit that researchers have to probe RNA-protein interactions, but these technologies are still in their infancy and thus new methods and applications are required to move our understanding forward.

We developed a novel, high-throughput approach to globally identify regions of RNAs that interact with proteins throughout a transcriptome of interest. We applied this technique to human HeLa cells and provide evidence that our approach captures both known and novel RNA-protein interaction sites. We identified global patterns of RNA-protein interactions, found evidence for co-binding of functionally related genes, and revealed that disease associated single-nucleotide polymorphisms are enriched within protein interaction sites.

We also performed detailed analysis of the RNA targets for two specific RBPs; Poly(A)-binding protein cytoplasmic 1 (PABPC1) and Argonaute (AGO). First, we used CLIP-seq to generate a transcriptome-wide map of PABPC1 interaction sites in the mouse transcriptome. This analysis revealed that PABPC1 binds directly to the highly conserved polyadenylation signal sequence and to translation initiation and termination sites. We also showed that PABPC1 binds

to A-rich regions in the 5’ untranslated region of a subset of messenger RNAs (mRNAs) and negatively regulates their gene expression.

Finally, we applied a recently developed approach to isolate and sequence AGO-bound microRNA precursors (pre-miRNAs). We uncovered widespread trimming and tailing, identified novel intermediates and created an index for pre-miRNA processing efficiency. We discovered that numerous pre-miRNA-like elements are embedded within mRNAs, but do not produce functional small RNAs. In total, these studies provide several advances in our understanding of the global landscape of RNA-protein interactions and serve as a foundation for future mechanistic studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: INTRODUCTION

This section refers to work in:

Silverman IM*, Li F*, Gregory BD. 2013. Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. *Plant Science*. 205-206:55-62

Abstract:

The eukaryotic transcriptome is regulated both transcriptionally and post-transcriptionally. Transcriptional control was the major focus of early research efforts, while more recently post-transcriptional mechanisms have gained recognition for their significant regulatory importance. At the heart of post-transcriptional regulatory pathways are *cis*- and *trans*-acting features and factors including RNA-binding proteins (RBPs) and their recognition sites on target RNAs. Recent advances in genomic methodologies have significantly improved our understanding of RBPs and their regulatory effects within the eukaryotic transcriptome. In this section, I will introduce these regulatory factors and describe the approaches for studying RNA-protein interaction sites, with an emphasis on recent methodological advances that produce transcriptome-wide datasets.

1.1 THE REGULATION OF GENE EXPRESSION

In the 1950's, Francis Crick proposed the central dogma of molecular biology, which in its simplest form states that genetic information flows from DNA to RNA through transcription, and from RNA to protein through translation (Figure 1.1) [1]. This elegant model serves as a basis for our understanding of molecular biology and gene expression in general.

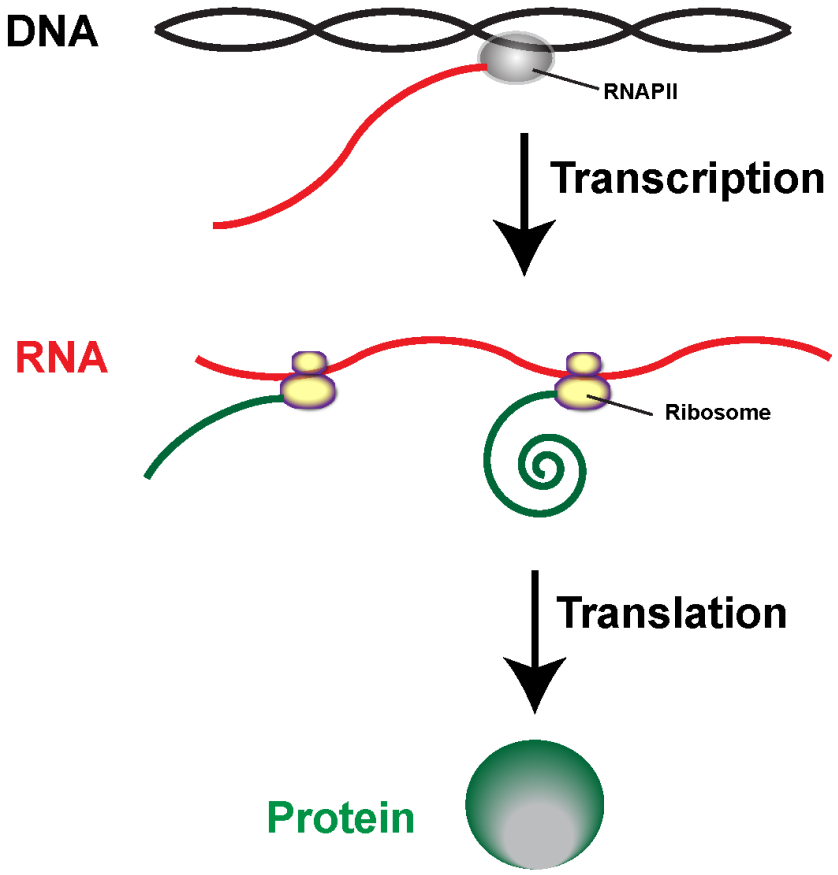


Figure 1.1 The central dogma of molecular biology. Genomic DNA is transcribed into mRNA by RNA polymerase II (RNAPII) and mRNA is translated into protein by the ribosome.

While Crick's model holds true for much of molecular biology, it gives us a static view of the complex and dynamic systems that are living biological organisms. How do organisms determine when to initiate and terminate transcription? Specifically, how do cells respond to changes in their environment? In 1961, Jacques Monod, discovered that *E. coli* lactose metabolism enzymes were only expressed in the presence of lactose and in the absence of glucose [2]. This seminal discovery of the lac operon was the first demonstration of transcriptional gene regulation and set a new paradigm in molecular biology.

Simple models of transcriptional gene regulation were sufficient to explain many observations in bacteria, but higher eukaryotes present a unique challenge to Monod's model.

How do organisms with billions of cells and hundreds of distinct cell types, each with highly specialized functions, regulate gene expression? The full answer to this question is outside the scope of this dissertation. Briefly stated, higher eukaryotes have evolved tens of thousands of proximal and hundreds of thousands of distal regulatory elements, which work in concert with regulatory proteins (transcription factors) to regulate the spatiotemporal expression of the approximately 20,000 protein coding genes encoded in mammalian genomes [3].

1.2 POST-TRANSCRIPTIONAL GENE REGULATION

Transcription of RNA is only the first process regulating gene expression (Figure 1.1). Once RNA is transcribed, numerous mechanisms exist that control the abundance, timing and even the sequence of proteins that are ultimately produced. Post-transcriptional regulatory processes allow cells to diversify their proteome, respond to environmental cues, and fine tune gene expression. This regulation can occur at any step of the RNA “life cycle” including maturation (e.g. 5’ capping, splicing, polyadenylation, etc.), transport from the nucleus, localization within subcellular compartments, molecule stability, as well as the initiation, elongation, and termination of protein translation (Figure 1.2). The integration of transcriptional and post-transcriptional processes ultimately determines the amount of each individual protein that is produced. Importantly, the stability and activity of proteins is subject to further regulation, but this is outside the scope of this discussion.

1.2.1 Alternative pre-mRNA splicing

Transcription results in the production of a pre-mRNA molecule that contains exons separated by long intervening sequences, called introns (Figure 1.2). In order for pre-mRNAs to mature into protein coding units, exons must be spliced together by the action of a multi-subunit macromolecular machine known as the spliceosome. Components of the spliceosome, including

the small nuclear ribonucleoproteins (snRNPs), recognize sequence elements in the exons and introns and catalyze the joining of exons to form mature mRNAs. However, it was observed in the late 1970's that the same pre-mRNA can give rise to multiple distinct isoforms, which are generated through alternative splicing reactions [4]. Depending on which exons are included and in which order, alternative mRNA isoforms can code for distinct proteins, contain regulatory sequences, or even contain premature stop codons leading to rapid decay [5]. This is one of the mechanisms by which higher eukaryotes diversify their limited set of 20,000 protein coding genes. We now understand that these alternative splicing events are mediated by specific regulatory sequences and structures in exons and introns, known as splicing enhancers and silencers, which interact with RBPs to promote or repress exon splicing. The complex rules which govern alternative splicing have only just begun to be elucidated [6, 7]

1.2.2 Alternative Polyadenylation

Polyadenylation of pre-mRNA represents another step by which the mature mRNA sequence can be altered. Polyadenylation of mRNAs is a key step in their maturation and is required for the transport, stability and productive translation of almost all mRNAs [8]. During transcription, RNA polymerase II (RNAPII) continues transcribing RNA through the end of the last exon. The polyadenylation machinery assembles on the 3' end of the pre-mRNA by interacting with specific sequence elements; most notably, the cleavage and polyadenylation specificity factor (CPSF), which binds to the polyadenylation signal (PAS; AAUAAA) approximately 20-25 nucleotides upstream of the eventual cleavage sites [9]. Another complex, the cleavage stimulation factor (CSTF), assembles downstream of the cleavage site and together with CPSF promotes cleavage, followed by subsequent polyadenylation of the mRNA. Consequently, a 5' to 3' exoribonuclease, Rat1, chases down the transcribing RNAPII and terminates transcription [10]. It later was noted that mRNAs contain multiple PAS sequences in their 3' UTRs, and more recent evidence suggests that ~75% of genes are subject to alternate polyadenylation (APA) [11]. If APA

occurs, the resulting mRNA sequences may vary not only in length, but also by the presence or absence of specific regulatory elements in the 3' UTR [8]. These regulatory elements may dictate the stability or translation efficiency, among other post-transcriptional processes.

1.2.3 Cytoplasmic Regulation of mRNA Stability and Translation

Once mature mRNA in exported into the cytoplasm, its lifespan and productivity are determined by the cohort of *cis*-regulatory elements and by the abundance of cognate *trans*-factors that it interacts with. Properly processed mRNAs will emerge from the nucleus carrying a protective 5'-7-methyguanlate (m^7G) cap and a 3' poly(A) tail. These features recruit protein factors, which aid generally in the stability and translation of the mRNA. However, a large amount of variation exists in both the stability and translation efficiency of mRNAs [12, 13]. For example the β-globin mRNA is much more stable than housekeeping mRNAs in erythrocytes [14, 15]. It is now well understood that RBPs and microRNAs interact with mRNAs through sequence and structure-specific interactions to regulate these two processes [15].

The mechanisms by which mRNA stability and translation are regulated are diverse. For example, in plants miRNAs generally cleave mRNA targets, leaving behind unprotected 5' and 3' ends that are rapidly degraded by the general degradation machinery [16]. In mammals, this mechanism is less often utilized, and rather miRNAs are thought to recruit deadenylation factors, which remove protective elements leading to decay [17]. Interestingly, many factors involved in mRNA turnover and miRNA mediated decay, accumulate in cytoplasmic processing bodies (p-bodies), possibly facilitating these functions (Figure 1.2) [18]. P-bodies are also thought to be a depot of transnationally repressed mRNAs, inhibiting translation by removing mRNAs from the translatable pool. Alternatively, numerous soluble factors can bind to the 5' UTR and inhibit ribosome scanning, which is a critical step in mRNA translation. While the processes I have described here are diverse, they are all controlled by a limited number of *cis*- and *trans*-acting elements. These include RBPs, microRNAs and their RNA recognition sites on mRNAs. We will

discuss these regulatory elements and the methods used to identify these regions in the transcriptome.

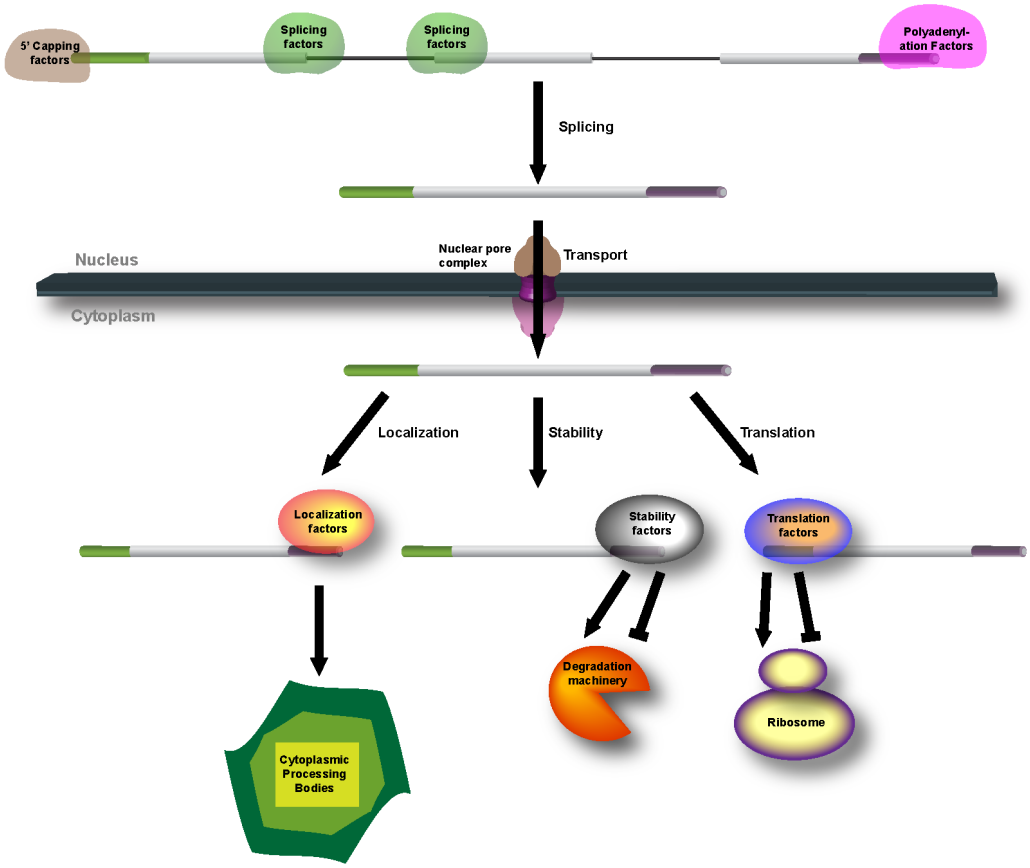


Figure 1.2 RNA-binding proteins mediate post-transcriptional gene regulation. RNAs are regulated by a variety of processes after transcription that are mediated by RBPs. White bars indicate coding exons. Green and purple bars indicate 5' UTR and 3' UTR, respectively.

1.3 RNA-BINDING PROTEINS

RNA-binding proteins (RBPs) are a group of *trans*-acting regulatory factors that are integral to the post-transcriptional regulation of eukaryotic transcriptomes. Cellular RNA is involved in a multitude of complex interactions with numerous RBPs from the initial processing of a transcript in the nucleus to its final translation and decay in the cytoplasm [19-21] (Figure 1.2). Recent experimental and bioinformatic analyses have suggested that >1,300 RBPs are encoded in the

human genome [22-24]. These proteins interact with mRNAs and form dynamic multi-component ribonucleoprotein (mRNP) complexes, which are the functional forms of mRNAs [25]. It is only through their proper formation that transcripts are correctly regulated and precisely produce the required amount of protein in a eukaryotic cell [19, 21, 25, 26]. Thus, RNA-protein interactions are necessary for the functionality, processing, and regulation of mRNA molecules.

1.3.1 RNA-Binding Domains

RBPs are a ubiquitous and heterogeneous class of proteins found in all organisms and characterized by the presence of one or more RNA-binding domains (RBDs). These proteins interact with single-stranded or double-stranded regions of RNA molecules through their binding domains, as well as with other cellular components through auxiliary domains. There are dozens of described RBDs, each with a distinct RNA-interaction interface. For instance, the RNA Recognition Motif (RRM) is the most abundant RNA-binding domain in mammalian cells (Table 1.1). The RRM is characterized by having a βαββαβ secondary structure, with the two α-helices packed against a 4-stranded β-pleated sheet. Canonically, the β-sheet is responsible for recognition of ssRNA (2-8 nucleotides), and the outward facing amino acid side chains in turn dictate the sequence specificity. The double-stranded RBD (dsRBD) is a common RBD that interacts with structured regions of RNA (Table 1.1). These RBDs are characterized by a ~65 amino acid domain in an αβββα structural arrangement in which the two α-helices overlap and pack against the antiparallel tri-β-sheet [27]. This structure allows the RBD to recognize the phosphate backbone and clamp onto a double-stranded RNA in lieu of a sequence motif. Other common RBDs include the K homology (KH) domain, cold-shock domain (CSD), several types of zinc finger (ZnF) domains (the most abundant being C-x8-X-x5-X-x3-H), DEAD/DEAH box, PIWI/Argonaute/Zwille (PAZ), and like-SM (LSM) (Table 1.1) [25, 28, 29]. Based on recent studies of RNA-interacting proteins, it is likely that many more RBDs are yet to be discovered [22-24].

1.3.2 Auxiliary Domains

RBPs are highly modular and may contain a single binding domain (e.g. DAZL), multiple copies of the same domain (e.g. PABPC1 and PCBP2), or a collection of different domains (e.g. IGFBP1) [24]. Together, the collection of RBDs in an RBP determines the affinity for target RNAs and increase specificity over a single RBD. Many RBPs possess auxiliary domains that carry out a variety of functions, such as facilitating protein-protein interactions or acting as substrates for post-translational modifications. Glycine-rich and arginine-serine-rich domains are common auxiliary domains observed in plants and metazoans [30, 31]. Auxiliary domains can have vast impacts on the mRNA target repertoire and regulatory potential of an RBP. For example a protein-protein interaction domain in GW182, a core component of the RNA induced silencing complex, interacts with Poly(A) binding proteins to recruit deadenylases to microRNA targeted mRNAs [32]. The presence or absence of a nuclear localization signal determines the subcellular localization and therefore the target RNA repertoire and functional outputs of an RBP. Therefore, the rules that govern RBP-RNA interactions are complex and understanding the *in vitro* specificities of an individual RBD does little to enhance our understanding of the true biological targets of a given RBP, *in vivo*.

Table 1.1 Known RNA-binding domains and the number of RBPs in humans containing these domains

RNA-binding domain	Human RBPs (Pfam)
RRM	597
KH	113
CSD	18
DS-RBD	50
ZnF (C-x8-C-x5-C-x3-H)	64
DEAD/DEAH box	200
PPR	8
RGG box	152
PUF	8
PAZ	12
LSM	35

1.3.3 Recurring Themes in RBP Biology

Detailed studies of RBPs have pointed to several recurring themes in RBP-mediated regulation. First, these proteins generally participate in multiple post-transcriptional processes, making the functional categorization of RBPs difficult. A prominent example of an RBP with multiple roles is SF2/ASF, which was originally identified as an essential splicing factor, and has now been implicated in translational control [33] and miRNA processing [34]. Second, a number of RBPs bind to and auto-regulate their own mRNAs, including DGCR8, RBFOX, TDP-43 and HuR [35-38]. Auto-regulation can occur via any of the mechanisms described in the previous section. Finally, mRNAs interact with multiple RBPs, which in turn bind to functionally related sets of mRNAs, suggesting a combinatorial network for control of gene expression at the RNA level [39]. Thus, the final fate of an mRNA is determined by the entire complement of bound RBPs. These themes point to a highly coordinated and controlled system of gene regulation by RBPs.

1.3.4 Poly(A)-Binding Proteins

The poly(A)-binding proteins (PABPs) are an important class of RBPs with global and gene-specific roles in regulating mRNA stability and translation efficiency. Canonically, PABPs exert their function by binding to the poly(A) tail, a post-transcriptional modification that is found on the 3' end of nearly all mRNAs (Figure 1.3A). Through this binding, PABPs are thought to physically protect the mRNA from 3' to 5' exonucleolytic decay and to interact with other *trans*-factors that bind to the mRNA 5' cap to promote translation [40]. Paradoxically, PABPs have also been shown to participate in negative regulation of mRNA stability through direct interaction with components of the RNA induced silencing complex (Figure 1.3B) [32]. More limited evidence suggests that PABPs interact with genomically encoded A- and AU-rich sequences in specific mRNAs to exert mRNA-specific regulation (Figure 1.3C) [41]. This mRNA-specific regulation can promote or repress translation, depending on the position of the binding [40]. Thus, PABPs can exert their function through a variety of mechanisms and through a number of *cis*- or *trans*-regulatory elements.

In mammals, there are six defined PABP isoforms; a single nuclear isoform, PABPN1, that impacts the addition of poly(A) tails in the nucleus and five cytoplasmic PABPs; ePAB, PABPC1, PABPC2, PABPC4, and PABPC5, that are thought to play roles in regulating mRNA stability and translation in the cytoplasm [42-44]. The overall structures and RNA binding specificities of the five cytoplasmic PABPs are highly conserved [45, 46]. They each contain four RNA Recognition Motifs (RRMs). RRMs 1 and 2 are primarily responsible for the high affinity binding to homopolymeric adenosines (K_d = 1.8 nM) [47], while RRMs 3 and 4 can bind to non-homopolymeric AU sequences (K_d= 2.9 nM) [47]. However, the levels of functional specificity and/or redundancy of the mammalian cytoplasmic PABPs remain unexplored.

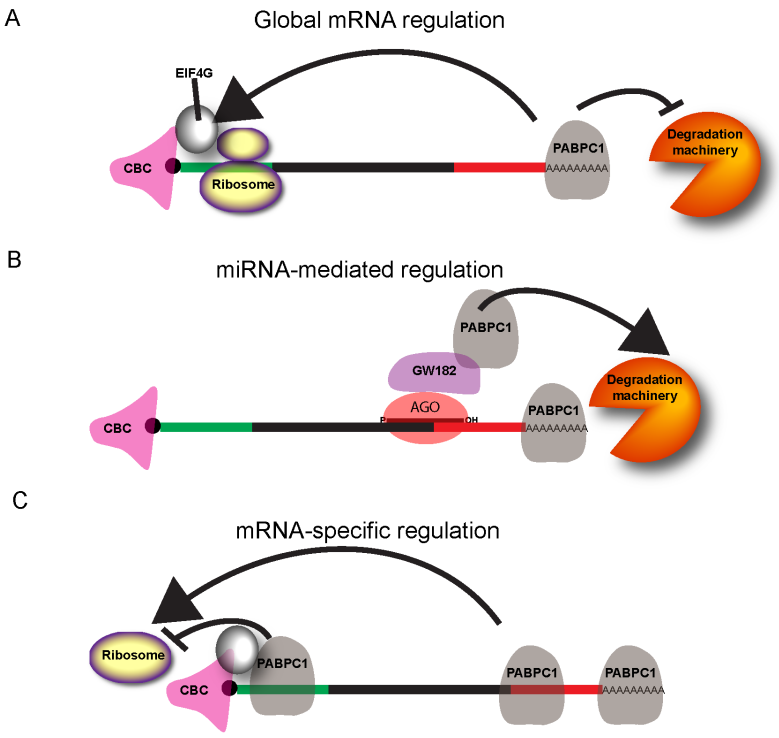


Figure 1.3 Mechanisms of PABPC1-mediated post-transcriptional gene regulation. A) PABPC1 plays a role in global mRNA regulation protecting mRNA from 3' end degradation factors and by interacting with EIF4G. B) PABPC1 participates in miRNA-mediated gene silencing by interacting with AGO proteins through GW182 to promote degradation. C) PABPC1 regulates specific mRNAs by disrupting ribosome scanning in the 5' UTR or promoting association with EIF4G from the 3' UTR.

PABPC1 is the major cytoplasmic PABP isoform in adult somatic cells and is abundantly expressed in all tissues [48]. The interaction of PABPC1 with mRNA poly(A) tails is well documented in multiple contexts [42, 49]. The corresponding functions of the PABPC1/poly(A) tail complex are primarily mediated in pathways of mRNA stabilization and translation enhancement (Figure 1.3A) [50-52]. These functions are linked to the interactions of PABPC1 with the 5' cap-binding complex (CBC) *via* heterodimerization with eIF4G [53, 54]. Through this interaction, PABPC1 is inferred to facilitate mRNA circularization, although this model has not been fully elucidated.

PABPC1 also plays a role in mRNA-specific gene regulation via two main mechanisms. PABPC1 interacts with GW182, which in turn interacts with Argonaute (AGO), the central mediator of RNA silencing [55]. Through this interaction, PABPC1 helps to recruit deadenylation and decay factors directly to the RNA, resulting in turnover (Figure 1.3B). Limited evidence also points to specific binding sites and functions for PABPC1 within genomically encoded regions of the eukaryotic mRNA transcriptome. For example, PABPC1 has been shown to bind to an A-rich element in the 5' untranslated region (UTR) of its own mRNA in mouse and human, and repress translation, establishing an auto-regulatory translational control circuit (Figure 1.3C) [41, 56, 57]. Due to the central role of PABPC1 in regulating global mRNA stability and translation, the impact of its auto-regulation on the transcriptome and proteome are vast. Analysis of the PABPs in the plant model system *Arabidopsis thaliana* suggests that this interaction is conserved in multiple organisms, representing an ancient RBP-mediated regulatory circuit [58]. The extent to which PABPC1 directly regulates other mRNAs in a similar fashion has not been explored.

A recent study in *Saccharomyces cerevisiae* using a photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation approach (PAR-CLIP) demonstrated *in vivo* binding of yeast poly(A) binding protein Pab1 to AU-rich elements in mRNAs [59], including binding to the efficiency element (UAUAUA) of the yeast polyadenylation signal [60, 61]. The downstream effects of Pab1 binding to the polyadenylation efficiency element in yeast remains undefined, as does any generalization of these findings to higher eukaryotic organisms. Based on its participation in global and mRNA-specific regulation through a variety of pathways, PABPC1 represents one of the most important RBPs in the mammalian genome. However, identification of PABPC1 targets in mammalian cells has not been performed to date.

1.4 MICRORNA BIOGENSIS

In addition to directly regulating mRNAs, RBPs also serve as biogenesis factors and effectors of microRNAs (miRNAs), another important class of post-transcriptional regulatory molecules. MiRNAs are short ~22 nucleotide small RNAs that function as sequence-specific guides to repress mRNA translation or stability. MiRNAs are conserved from plants to mammals; however distinct biogenesis pathways suggest that the miRNA system evolved at least twice [16]. The human genome encodes thousands of miRNAs, each of which can bind to and regulate hundreds of mRNAs [62]. In humans, specific miRNAs have been implicated in numerous biological pathways, are misregulated in disease, and have conserved functional roles in eukaryotes [63-65]. Thus, understanding the biogenesis, regulation and function of these small RNA molecules is critical to our understanding of gene regulation.

1.4.1 Nuclear Processing

In mammals, miRNAs are transcribed by RNA polymerase II (RNAPII) as primary miRNA (pri-miRNA) genes or as pieces of larger parent RNA molecules. In fact, the majority of human miRNAs reside within introns, with only a handful of miRNAs identified within mRNA exons [66]. Regardless of their origin, miRNA stem-loops are processed into miRNA precursors (pre-miRNAs) by the microprocessor complex, which is comprised of the type III ribonuclease Drosha and the dsRBD-containing RBP, DGCR8 [67, 68] (Figure 1.4). The microprocessor complex binds to stem-loop structures in the nucleus and cleaves a ~65 nucleotide (nt) pre-miRNA molecule ~11 nucleotides from the base of the stem with a 2nt 3' overhang, which enhances Dicer processing [69]. Recent studies have found that pre-miRNA biogenesis can occur in a microprocessor-independent fashion whereby pre-miRNAs are directly generated by the spliceosome [70, 71]. After these initial processing steps, pre-miRNAs are transported into the cytoplasm by the nuclear transport protein Exportin-5 to be further processed [72, 73].

1.4.2 Cytoplasmic Processing

In the cytoplasm, pre-miRNAs interact with the miRNA loading complex (miRLC), which consists of the miRNA effector protein Argonaute (AGO), another Type III endonuclease called Dicer, and dsRBD-containing TRBP [74-76] (Figure 1.4). Dicer is responsible for cleaving the pre-miRNA on the stem-loop side of the duplex, leaving a ~22nt miRNA-miRNA* duplex, with 2 nt 3' overhangs on both ends. Dicer is able to cleave pre-miRNAs in the absence of AGO and TRBP. However, recent evidence suggests that the miRLC is the major pre-miRNA maturation pathway in mammals *in vivo* [77]. Interestingly, Dicer-independent pathways have also been discovered for mammalian miRNA maturation [78]. For instance, the erythrocyte specific pre-miR-451 contains a short stem loop that is a poor substrate for Dicer processing. Instead, AGO2, which is the only AGO in mammals with catalytic activity for RNA cleavage, cuts pre-miR-451 as part of the miRNA precursor deposit complex (miPDC), promoting 3' to 5' exonucleolytic decay by the poly(A) ribonuclease, PARN, resulting in a mature and active miR-451 [79, 80] (Figure 1.4). AGO2-cleaved pre-miRNAs (ac-pre-miRNAs) function as an alternative biogenesis mechanism for several other pre-miRNAs, although the extent to which pre-miRNAs can be processed along this pathway has not been addressed [78].

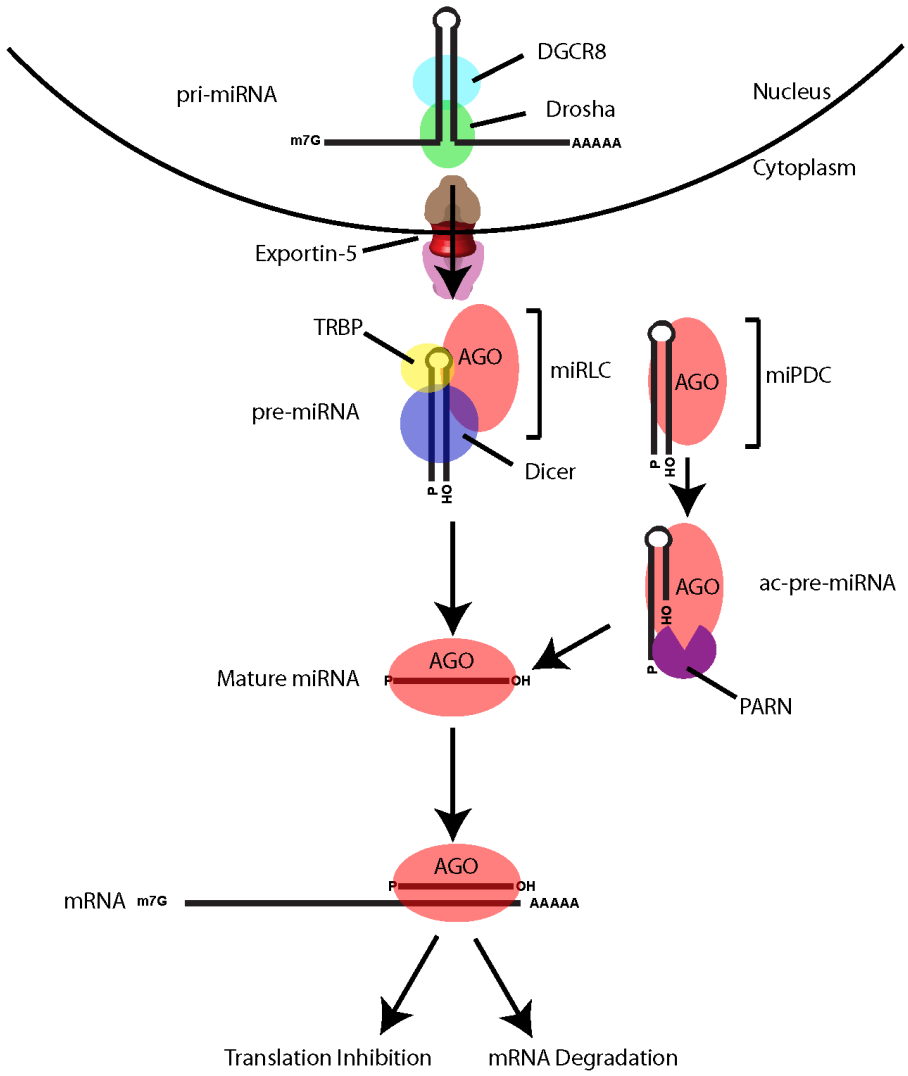


Figure 1.4 Mammalian microRNA biogenesis and function. Pre-miRNAs are processed from primary miRNAs (pri-miRNA) or from other host RNA species in the nucleus by the DGCR8/Drosha microprocessor complex. Once in the cytoplasm the pre-miRNAs interact with the miRLC to be processed by DICER. Alternatively, some pre-miRNAs (ac-pre-miRNAs), are cleaved by AGO2 and trimmed by PARN to produce functional miRNA (miPDC). Only one strand is loaded and used as a guide for RNA silencing.

1.4.3 AGO-Loading and miRNA Function

Canonically, one of the strands in the liberated miRNA-miRNA* duplex is selectively loaded into one of four AGO proteins to make a functional RNA-induced silencing complex

(RISC) (Figure 1.4). However, there is some evidence that the miRNA* (passenger strand) may serve other functions in the cell and miRNA strand switching has been observed between cell types [81, 82]. miRNA-RISC binds to target mRNAs through complementary base-pairing interactions, which in mammals are primarily dependent on the miRNA seed sequence (nucleotides 2-8 from the 5' end of the miRNA) [83]. More recent evidence has demonstrated that sequences outside the seed are important for miRNA target recognition, and that *in vivo* miRNA-mRNA target pairs do not always follow seed pairing rules [84, 85].

It is clear that miRNAs negatively regulate their target mRNAs, although the exact mechanisms by which they exert their regulatory function remains controversial [16] (Figure 1.4). Evidence suggests that some combination of translation inhibition and mRNA degradation contribute to the decreased mRNA and protein abundance of miRNA target mRNAs [86-88]. The exact means of miRNA-mediated repression may depend on numerous factors, including sequence complementarity, binding site accessibility, and the presence of other factors including specific RBPs.

1.5 METHODS TO STUDY RNA-PROTEIN INTERACTIONS

How do RBPs recognize their RNA targets? As described earlier, the β-sheet in each RRM is responsible for recognition of a specific sequence element. However, given the variety of RBDs, and the multitude of these domains in each RBP, recognition of RNA targets is governed by complex rules. Further complications arise when one considers that some such domains (i.e. dsRBD) do not recognize specific sequence elements but rather specific RNA structures. A prime example of this is the recognition of miRNA stem loop by DGCR8. This protein binds to a specific structural arrangement found in thousands of miRNA stem-loops rather than a clearly defined sequence motif [89]. In actuality, both the primary sequence and secondary structure of RNA targets are important for target recognition. Therefore, multiple approaches are required to gain an understanding of the features that dictate RBP-RNA interactions. Here, I will review classical

approaches and discuss more recent methodologies that directly identify RNA-protein interaction sites in cells.

1.5.1 Classical Approaches

A comprehensive analysis of bound RNA targets is necessary to understand the role of RBPs in post-transcriptional gene regulation. This information is needed to determine the specific binding sites as well as the sequence and structural preferences (interaction motif(s)) of each RBP. Initially, *in vitro* approaches were developed to identify these interacting motifs. Such approaches include RNA Electrophoretic Mobility Shift Assays (EMSAs), RNA-affinity chromatography, UV-crosslinking studies, Systematic Evolution of Ligands by Exponential Enrichment (SELEX), and RNACompete [90-95]. Although these studies have proven useful in identifying RBP interacting motifs and *cis*-elements, they are performed *in vitro* and thus may not reflect biologically relevant sequence specificities in cells.

RNA EMSAs utilize *in vitro* binding and non-denaturing gel electrophoresis to identify changes in gel mobility due to binding events of protein and nucleic acids [92]. While effective in demonstrating strong protein-nucleic acid interactions (especially DNA-protein), EMSAs may not be sensitive enough to capture weak or transient binding events. UV-crosslinking experiments, in which covalently linked RNA-protein complexes are interrogated by SDS polyacrylamide gel electrophoresis (SDS-PAGE), can be utilized to increase sensitivity [91]. In RNA-affinity chromatography, a specific RNA sequence is used to capture an interacting RBP(s) from a total protein cell lysate [90]. This approach is commonly used when trying to identify protein partners of known *cis*-regulatory sequences, but highly abundant or promiscuous RBPs may confound results. Conversely, SELEX provides an approach to identify specific protein-interacting sequences for a particular protein of interest [93]. SELEX reduces investigator bias but systematic biases may also exist due to the *in vitro* nature of the methodology. Recent advances in SELEX-like approaches have enabled more high-throughput analyses and generated a

valuable resource of RNA-binding site sequence preferences for a few hundred RBPs [94, 95]. While all of these approaches can reveal RNA-protein interactions, each method has disadvantages and the most reliable results are those confirmed by multiple methods. Furthermore, most of these approaches can only be performed on one RNA or protein at a time, and may not consider RNA secondary structure, severely limiting their usefulness.

1.5.2 Genomic Era Approaches

More recently, *in vivo* approaches have been developed to directly study RNA-protein interactions in cells (Figure 1.5). All of these methods rely on the same general scheme, whereby RBPs are co-immunoprecipitated with their RNA targets, followed by identification and quantitation of bound RNAs. For instance, RNA immunoprecipitation (RIP) followed by RT-PCR, microarray (RIP-chip), or high-throughput sequencing (RIP-seq) have been used extensively to identify mRNA targets of RBPs from a variety of organisms [96]. RIP can also be performed in the presence of formaldehyde to stabilize interactions between RNAs and their interacting proteins. This method allows for more stringent washing and reduces the levels of RBP association with non-biologically relevant targets after cell lysis [97]. One caveat of this approach is that formaldehyde also crosslinks proteins to one another, and therefore the identified interactions may be indirect. However, revealing indirect associations may also be informative and biologically relevant given the complex nature of mRNPs in eukaryotic cells.

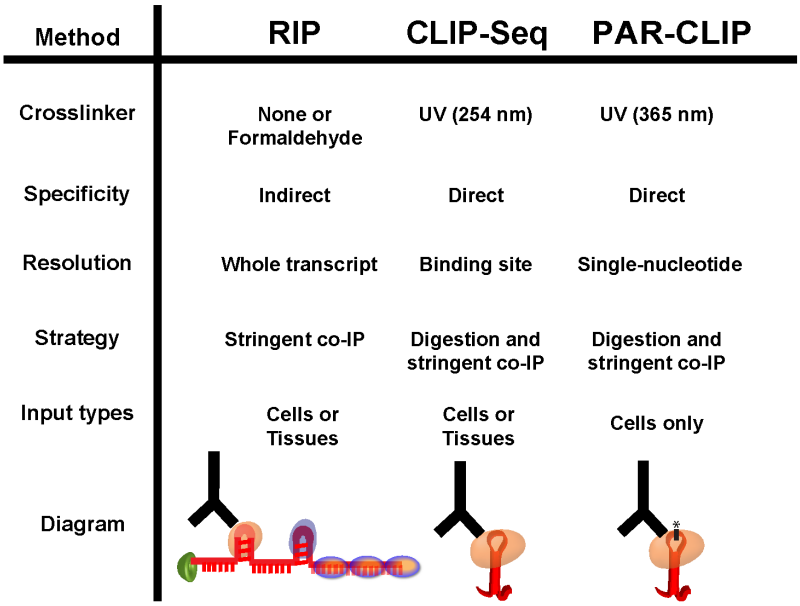


Figure 1.5 Global approaches to Identify RNA-protein interaction sites. In RIP, whole mRNAs are immunoprecipitated and quantified by qPCR, microarray, or sequencing. In CLIP-seq, RNA fragments are immunoprecipitated and sequenced to identify clusters. In PAR-CLIP RNA fragments are immunoprecipitated and sequenced and T>C transversions are used to identify single-nucleotide binding sites.

A more specific approach for defining RNA-protein interactions is the Crosslinking and Immunopreciptiation (CLIP) approach (Figure 1.6 and 1.7). This approach relies on the crosslinking specificity of UV (254 nm) light, which covalently attaches RNAs to their interacting proteins (Figure 1.5) [98]. A ribonuclease (RNase) digestion is performed during the isolation of the RNA-protein complexes, thereby revealing the specific interaction regions of RNA targets (Figure 1.6). This improves the resolution of CLIP by isolating only RBP interacting sites in contrast to the full length RNA molecule that is isolated in RIP-based studies (Figure 1.5). CLIP followed by high-throughput sequencing-based analysis of protein-bound RNA sites (CLIP-seq or HITS-CLIP) and several variant protocols (e.g. Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP)) have been widely used to study RBPs in a diverse set of metazoan cell types (Figure 1.6) [99-101]. In PAR-CLIP, 4-thiouridine is introduced

into the cell media and longer wave UV light (365 nm) is used to specifically crosslink this non-natural nucleotide. Cross-linking events create transversion mutations (T(U)>C) and algorithms are subsequently used on the resulting sequencing data to identify single-nucleotide resolution binding sites. These methods reveal the entire complement of binding sites for a given RBP, and have provided enormous insight into the role of these proteins in pre-mRNA splicing [102, 103], stability [104], and translation [105].

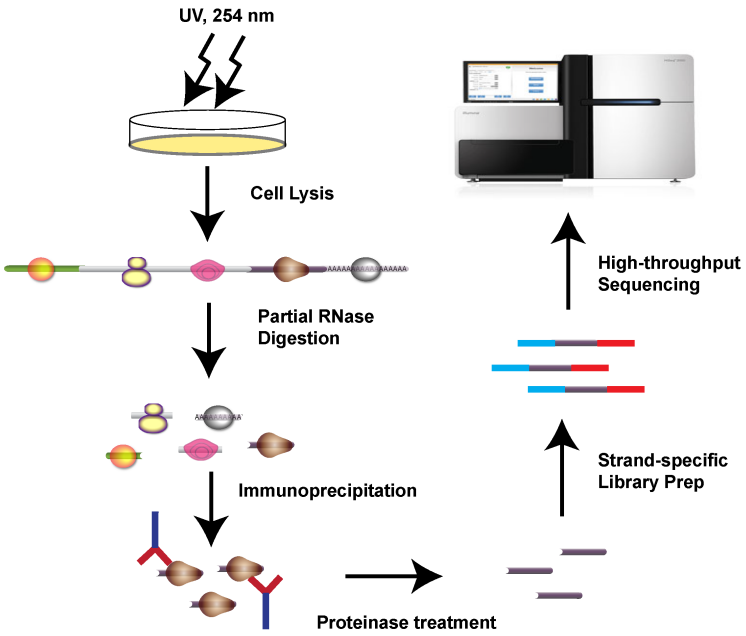


Figure 1.6 Overview of the CLIP-seq approach. In CLIP-seq, RNA-protein complexes are crosslinked with UV light (254 nm). RNAs are digested through an RNase treatment and a protein of interest is immunoprecipitated. Crosslinks are reversed by proteinase digestion followed by strand-specific library preparation and high-throughput sequencing.

While CLIP-seq and PAR-CLIP are powerful approaches to identify *in vivo* RNA-protein interaction sites, they can only identify the binding sites of a single protein at a time. Therefore, their impact is limited due to the large number of RBPs encoded in genomes as well as the labor-intensive nature of these assays. Therefore, there is a need for more global approaches for defining RNA-protein interaction sites. The work discussed in Chapter 2 describes the development of one such assay by our laboratory (Figure 1.7). Concurrent with this work, other

RNA-centric approaches were developed. For instance, a photoactivatable-ribonucleoside enhanced crosslinking (PAR-CL) and oligo-dT affinity purification coupled with RNase (RNase I) digestion was used to comprehensively reveal the binding sites of RBPs along mature mRNAs in human and yeast (Figure 1.8) [22, 106]. These approaches are an RNA-centric means to define RNA-protein interaction sites across eukaryotic transcriptomes without the need for antibodies to specific proteins. We and others have since used these approaches to investigate nuclear RNPs and compare RBP-RNA interaction profiles in different cell types [107, 108]. Future studies are necessary to address RNA-protein interaction dynamics during important biological processes in order to advance our understanding of post-transcriptional gene regulation.

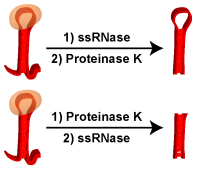
Method	PAR-CL	PIP-seq
Crosslinker	UV (365 nm)	UV (254 nm) or Formaldehyde
Specificity	Direct	Direct or Indirect
Resolution	Single-nucleo-tide	Binding site
Strategy	Crosslinking and digestion	Crosslinking and differential digestion
Input types	Cells only	Cells or Tissues
Diagram		
Examples	HEK293T, HeLa, Yeast	HeLa, HEK293T, A. Thaliana

Figure 1.7 RNA-centric approaches to study RBP-RNA interactions. In PAR-CL, cells are supplemented with 4-thiouridine, and crosslinked with UV (365 nm). mRNA-protein complexes are enriched by oligo-dT selection and RNase/proteinase digestion is performed to liberate complexes. In PIP-seq, RNA-protein complexes are stabilized with formaldehyde followed by differential RNase digestion. Protein binding sites are identified by comparative analysis.

1.6 OUTLINE OF DISSERTATION

In Chapter 2, I describe the development of a novel, high-throughput approach to globally identify regions of RNA-protein interaction throughout a transcriptome of interest. As a proof-of-principle, we applied this technique to human HeLa cells and provide evidence that our approach captures both known and novel RNA-protein interaction sites. We identified global patterns of RNA-protein interactions, found evidence for co-binding of functionally regulated genes, and revealed that disease associated single nucleotide polymorphisms (SNPs) are enriched within RBP interaction sites.

In Chapter 3, we used CLIP-seq to create a transcriptome-wide map of poly(A)-binding protein (PABPC1) interaction sites in the mouse transcriptome. This analysis revealed that PABPC1 binds to mammalian mRNAs outside of its canonical role in poly(A) tail binding. We showed that PABPC1 binds directly to the highly conserved polyadenylation signal sequence and to translation initiation and termination sites. We also showed that PABPC1 binds to A-rich regions in the 5' untranslated region of a subset of mRNAs, including its cognate mRNA, and negatively regulates their translation and stability.

In Chapter 4, we applied a recently developed approach to isolate and sequence Ago-bound pre-miRNAs in the human transcriptome. Using a novel bioinformatics pipeline, we uncovered widespread trimming and tailing of pre-miRNAs and identified novel AGO2-cleaved pre-miRNAs. We created an index for pre-miRNA processing efficiency and discovered that numerous pre-miRNA-like elements are embedded within mRNAs. Some of these represent novel miRNAs but the majority, including the iron-responsive element of ferritin genes, are inefficiently processed into mature small RNAs. The function of these poorly processed pre-miRNA-like sequences will be the focus of future investigations.

In Chapter 5, I discuss the implications of these studies, and delineate future experiments to address new questions that have arisen from this work.

Chapter 2: RNASE-MEDIATED PROTEIN FOOTPRINT SEQUENICNG

This section refers to work from:

- Silverman IM*, Li F*, Alexander A, Goff L, Cole T, Rinn JL, Gregory BD. 2014. RNase-mediated protein footprinting reveals protein-binding sites throughout the human transcriptome. *Genome Biology.* 15:R3
- Silverman IM and Gregory BD. Transcriptome-wide Ribonuclease footprinting to identify RNA-protein interaction sites. *Methods*. 72:76-85

Abstract:

RNAs are continuously associated with RNA-binding proteins (RBPs), and these interactions are necessary for many key cellular processes ranging from splicing to chromatin regulation. Although numerous approaches have been developed to map RNA-binding sites of individual RBPs, few methods exist that allow assessment of global RBP-RNA interactions. Here, we describe a universal, high-throughput, ribonuclease-mediated protein footprint sequencing approach that reveals RNA-protein interaction sites throughout a transcriptome of interest. We apply this method to the HeLa transcriptome and compare RBP binding sites found using different cross-linkers and ribonucleases. From this analysis, we identify numerous putative RBP binding motifs, reveal novel insights into co-binding by RBPs, and uncover a significant enrichment for disease-associated polymorphisms within RBP interaction sites.

Contributions:

The contents of this section were generated by in collaboration with Fan Li. I performed all experimental analyses with technical assistance from Anissa Alexander. Fan Li provided bioinformatic support for the computational aspects of the work and assisted in the drafting of the first manuscript.

2.1 INTRODUCTION

RNA-protein interactions are central to all of the post-transcriptional regulatory processes that control gene expression. From the initial processing of a protein-coding transcript in the nucleus to its final translation and decay in the cytoplasm, cellular mRNAs are involved in a complex choreography with various *trans*-acting RNA-binding proteins (RBPs) [19-21]. RBPs are also required for the processing and function of the thousands of non-coding RNAs (ncRNAs), both large and small, encoded by eukaryotic genomes. These RNAs have a variety of cellular functions, including chromatin regulation and control of cell fate [109, 110]. Thus, RNA-protein interactions represent a vast, diverse, and critical layer of transcriptome regulation.

Eukaryotic genomes encode a large collection of RBPs that interact with mRNAs to form dynamic multi-component ribonucleoprotein complexes (RNPs) [111, 112]. These RNPs often constitute the functional forms of mRNAs, and it is only through their proper formation that transcripts are correctly regulated to precisely produce the required amounts of each protein in a cell [19, 21, 26, 112]. Intriguingly, recent evidence suggests that post-transcriptional regulation of mRNAs encoding functionally related proteins likely requires mRNP assembly by specific sets of co-occurring RBPs, an idea that was originally postulated by the post-transcriptional operon hypothesis [39, 113]. Thus, the precise composition and formation of RNPs in eukaryotic cells is critical for proper gene expression regulation.

The essential nature of RNA-protein interactions to eukaryotic biology has led to the use of numerous biochemical, genetic, and computational approaches being utilized, alone and in combination, to identify and validate RBPs and their specific RNA-binding sites [20, 114, 115]. These approaches have proven useful in characterizing a number of RBPs [89, 103, 105, 116-126]. However, all of these earlier approaches investigated RNA-protein interactions one protein at a time, limiting the ability to monitor the global landscape of RNPs and to reveal insights into the combinatorial binding and regulation by the cellular milieu of RBPs. This observation points to a major gap between the significance of cellular RNA-RBP interactions and the difficulty in

establishing a comprehensive catalogue of these interactions in a single experiment.

Recently, several groups have established experimental approaches to interrogate RNA-protein interaction sites on a more global scale. These approaches utilize 4-thiouridine and ultraviolet (UV) cross-linking to identify RNA-protein interactions by uncovering sites of T>C transversion (representing RNA-protein cross-linking events) [127, 128]. However, these studies have been limited by several factors. Specifically, they rely on treatment with synthetic nucleotides and UV cross-linking, which can be used for cell culture but not tissues or whole organisms. Furthermore, UV cross-linking only identifies sites of direct RNA-protein contact and may not capture the larger multi-protein complexes that comprise the overall RNP architecture *in vivo*. Finally, these studies have focused on poly-adenylated (polyA) transcripts, reducing their ability to monitor RBP binding in non-polyA and nascent RNAs.

To address the limitations of the currently available methodologies, we report here a ribonuclease (RNase)-mediated protein footprint sequencing approach that we call protein interaction profile sequencing (PIP-seq). This approach identifies RNA-protein interaction sites within both unprocessed and mature RNAs in a mostly unbiased manner and on a transcriptome-wide scale. We describe the use of multiple cross-linking techniques to capture both direct and indirect RNA-protein interactions. We also show that both single-stranded and double-stranded RNases uncover distinct but overlapping sets of RNA-protein interaction sites. Using this approach, we find PIP-seq to be a reproducible approach that reveals both previously known and novel RBP interaction sites. We demonstrate the utility of PIP-seq by uncovering enriched sequence motifs within the complement of identified RBP interaction sites. We also investigate the interactions among protein-binding sites and provide evidence for co-binding of RNAs by specific sets of RBPs, some of which bind to groups of transcripts encoding functionally related proteins. These results reveal novel insights into networks of post-transcriptional gene regulation mediated by specific groups of RBP-bound sequence motifs. Finally, we identify a significant enrichment for disease-associated variants within RBP interaction sites, and demonstrate the

effects of some of these single nucleotide polymorphisms (SNPs) on RNA-protein interactions. Overall, our approach provides an RNA-centric global assessment of RNA-RBP interactions that directly identifies RNA-protein interaction sites and is applicable for use in all organisms and sample types.

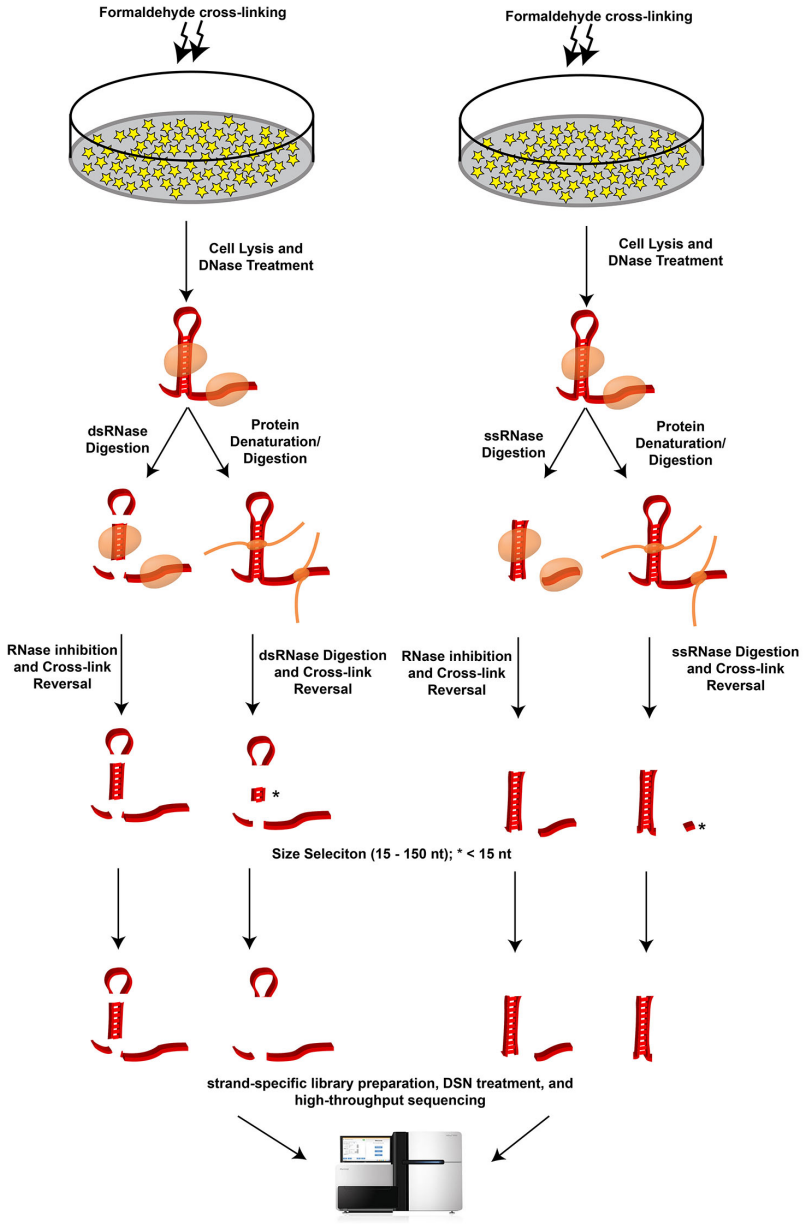
2.2 RESULTS AND DISCUSSION

2.2.1 RNase-mediated protein footprinting identifies sites of RNA-protein interaction

To obtain an unbiased, genome-wide view of RNA-protein interactions for both unprocessed and mature RNAs in eukaryotic transcriptomes, we developed an RNase-mediated protein footprint sequencing approach, referred to herein as PIP-seq, by performing our nuclease-sensitivity sequencing assays [129, 130] on cross-linked RNA-protein complexes from HeLa cells (Figure 1A). Previous investigations of RNA-protein interactions have assayed stable endogenous interactions as well as those captured by the use of UV (254 nm), which cross-links only direct protein-nucleic acid contacts, and formaldehyde, which cross-links protein-nucleic acid and protein-protein contacts with longer range [131-133]. Therefore, to generate a comprehensive and multifaceted view of RBP interaction sites, we used both cross-linking techniques and no cross-linking when performing PIP-seq.

We had previously used nuclease-sensitivity sequencing assays on purified RNAs to determine RNA base-pairing probabilities by treating RNA with either single-stranded or double-stranded RNase (ss- or dsRNase, respectively) and sequencing the resulting populations. We reasoned that by using both of these RNases on cross-linked RNA-protein complexes, we would be able to both comprehensively map RBP binding sites and also investigate RNA base-pairing probabilities *in vivo.* However, for the purposes of this manuscript we focus our analysis specifically on the identification of protein-interaction sites, which we refer to as Protein-Protected Sites (PPSs).

To perform PIP-seq, we started with adherent HeLa cells cross-linked by one of the methods described above (UV or formaldehyde) or used cells that had not been cross-linked. The resulting cell lysates were then split into experimental and background samples. Due to the structure-specific nature of the RNases used, it was essential to have a background sample to control for RNase insensitive regions. Therefore, a 'footprint sample' (experimental) was directly treated with either a single- or double-stranded RNase (ssRNase [RNaseONE] or dsRNase [Rnase V1], respectively). In contrast, the 'RNase digestion control' sample was first denatured in SDS and treated with Proteinase K prior to RNase digestion. In this way, regions that were protein-protected in the footprinting sample became sensitive to RNase digestion in the control sample and regions that were unbound but insensitive to one of the nucleases due to their structural status, remained that way. For both samples, cross-links were subsequently reversed (heating for formaldehyde cross-links and extensive Proteinase K treatment for UV cross-links) and followed by strand-specific library preparation (Figure 2.1). Highly abundant RNA species (e.g. ribosomal RNAs) were depleted from each library based on their rapid re-annealing rates using a duplex-specific thermostable nuclease (DSN) protocol (see Materials and Methods for more details).


Formaldehyde cross-linking
Formaldehyde cross-linking
Cell Lysis and DNase Treatment
Cell Lysis and DNase Treatment
dsRNase Digestion
Protein Denaturation/ Digestion
ssRNase Digestion
Protein Denaturation/ Digestion
RNase inhibition and Cross-link Reversal
dsRNase Digestion and Cross-link Reversal
RNase inhibition and Cross-link Reversal
ssRNase Digestion and Cross-link Reversal
Size Seleciton (15 - 150 nt); * < 15 nt
strand-specific library preparation, DSN treatment, and high-throughput sequencing

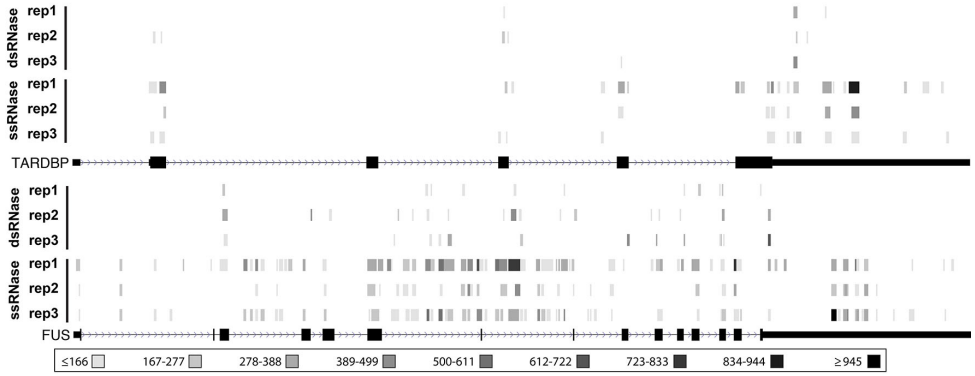

dsRNase
ssRNase
rep1
rep2
rep3
TARDBP
FUS
≤166
167-277
278-388
389-499
500-611
612-722
723-833
834-944
≥945

Figure 2.1 PIP-seq strategy and design. Tissue culture cells are cross-linked with formaldehyde and split into two samples. RNase footprinting samples are subjected to RNase treatment with either an ssRNase (RNase One) or dsRNase (RNase V1). RNases are then inhibited and cross-links reversed. RNase digestion control samples are subjected to protein denaturation and digestion first, followed by RNase treatment (ssRNase or dsRNase). The RNA fragments are then ligated between RNA sequencing adapters and subjected to strand-specific library preparation. DSN treatment is used to remove highly abundant RNA species and the resulting library is sequenced on an Illumina HiSeq2000. Examples of PPSs identified in TARDBP (top panel) and FUS (bottom panel) by replicates from ssRNase and dsRNase PIP-seq experiments. Color scale indicates CSAR enrichment score for footprint library compared to digestion control library (as indicated at the bottom of the figure).

We then sequenced the resulting libraries (4 total for each replicate) using the Illumina 50 base pair (bp) single-end sequencing protocol, and obtained ~31-60 million raw reads per library (Table 2.1). To identify PPSs, we used a Poisson distribution model based on a modified version of the CSAR software package [134]. Specifically, read coverage was calculated for each base position in the genome and a Poisson test was used to compute an enrichment score for footprint versus RNase digestion control libraries (Table 2.1). PPSs were then called as described for ChIP-seq analysis [134] with a false discovery rate (FDR) of 5% (Figure 2.1). Using this approach we identified a total of ~1,011,000 PPSs over 7 experiments, comprising ~430,000 non-overlapping sites (Table 2.1).

We found PPSs identified by both cross-linking strategies and with no cross-linking to be widely distributed across both exonic and intronic regions, with a particular enrichment for distal intronic binding in the formaldehyde cross-linked experiments (Figures 2.2). Closer examination of PPSs broken down by genic features (e.g. 5' and 3' UTR, CDS, and intron) or RNA type (mRNA and lncRNA) revealed that > 50% of all human mRNAs contained multiple binding events across all transcript regions except the 5' UTR (average of ~1 PPS in only 28.8% of total transcripts) in HeLa cells (Figure 2.3). Strikingly, an average of ~26 PPSs were found in the introns of each transcript in the formaldehyde cross-linked PIP-seq experiments, compared with ~3 and ~2 intronic PPSs with the UV and non-cross-linked experiments, respectively (Figure 2.3). These results suggest that formaldehyde cross-linking captures more transient and/or weak RBP-RNA interactions within intronic (especially distal (> 500 nucleotides (nt) from a splice site))

portions of mRNAs. We also found that ~2 – 6% of all known human lncRNAs could be identified as containing an average of 2.5 PPSs in HeLa cells using PIP-seq with the various cross-linking strategies (Figure 2.3). The limited number of PPS-containing lncRNAs uncovered by our experiments is likely due to the low expression and tissue-specific nature of these transcripts. To address a possible dependence of our approach on RNA expression levels, we assessed the relationship between RNA steady state abundance and number of PPSs per transcript and found that RNA levels explained only a small fraction (R^2 = 0.11) of the total variation in PPS counts between transcripts (Figure 2.4). Overall, these results suggest that PIP-seq provides a comprehensive and mostly unbiased view of global RNA-protein interaction sites in eukaryotic transcriptomes.

Table 2.1 PIP-seq library characteristics

Cross-linker	Rnase	Replicate	Library Type	raw reads	trimmed reads	trimmed reads (%)	mapped reads	mapped reads (%)	PPS (FDR=5%)
Formaldehyde	double-stranded	1	Footprint	60,880,156	42,030,874	69.04%	30,639,721	72.90%	70,371
			Control	77,929,058	65,922,052	84.59%	53,512,731	81.18%	
		2	Footprint	103,702,805	89,989,687	86.78%	74,858,669	83.19%	88,060
			Control	88,842,812	75,858,871	85.39%	67,295,403	88.71%	
		3	Footprint	66,398,039	59,750,511	89.99%	51,968,442	86.98%	122,277
			Control	77,342,721	59,282,909	76.65%	52,909,099	89.25%	
	single-stranded	1	Footprint	70,747,816	51,183,479	72.35%	45,281,100	88.47%	190,654
			Control	67,705,765	40,019,397	59.11%	36,112,186	90.24%	
		2	Footprint	70,546,971	58,144,499	82.42%	51,232,318	88.11%	289,984
			Control	62,663,571	48,642,460	77.62%	45,222,199	92.97%	
		3	Footprint	64,725,704	46,067,911	71.17%	40,107,905	87.06%	143,631
			Control	79,466,145	60,612,150	76.27%	56,223,519	92.76%	
UV (254 nm)	double-stranded	1	Footprint	31,019,360	27,834,338	89.7%	25,654,498	92.2%	6,642
			Control	39,136,707	35,069,030	89.6%	30,905,190	88.1%	
		2	Footprint	24,604,010	21,458,724	87.2%	18,832,305	87.8%	2,871
			Control	32,977,185	29,400,832	89.2%	22,126,579	75.3%	
	single-stranded	1	Footprint	31,248,062	25,230,672	80.7%	23,381,804	92.7%	42,878
			Control	29,411,398	24,114,686	82.0%	22,983,479	95.3%	
		2	Footprint	30,371,337	25,984,739	85.6%	24,337,412	93.7%	24,635
			Control	27,442,306	21,546,936	78.5%	20,412,845	94.7%	
None	double-stranded	1	Footprint	33,186,168	31,303,968	94.3%	29,057,193	92.8%	2,428
			Control	34,912,635	32,291,230	92.5%	27,402,521	84.9%	
	single-stranded	1	Footprint	32,691,777	28,246,801	86.4%	26,881,380	95.2%	26,594
			Control	29,148,805	24,234,319	83.1%	23,072,655	95.2%	
			Total	1,267,101,313	1,024,221,075		900,411,153		1,011,025
			Average	52,795,888	42,675,878	83%	37,517,131	89%	42,126

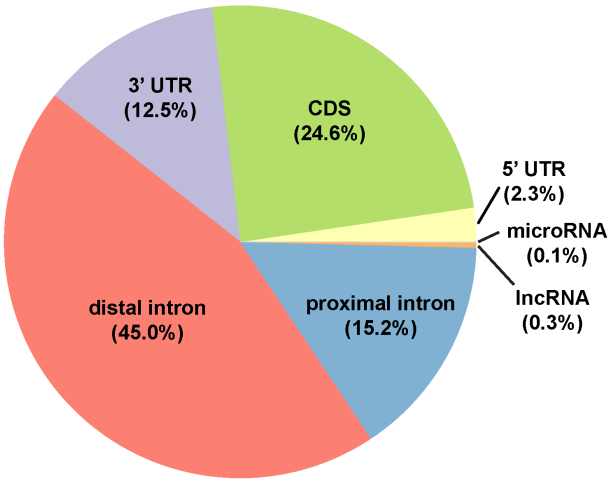


Figure 2.2 Absolute distribution of PPSs throughout RNA species for formladehyde PIP-seq experiments.

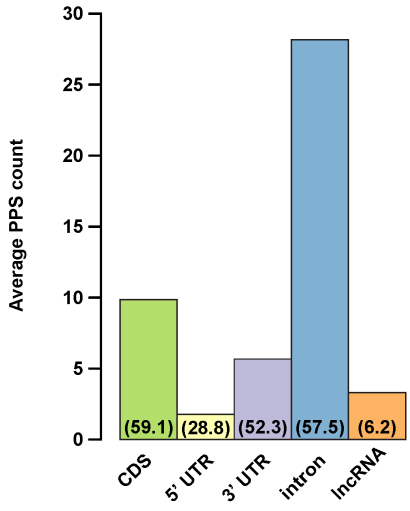


Figure 2.3 Average PPS count per RNA molecule (classified by RNA type (mRNA and lncRNA) and transcript region (e.g. 5' UTR)) for formaldehyde PIP-seq experiments. Percentages indicate the fraction of each RNA type or region that contains PPS information.

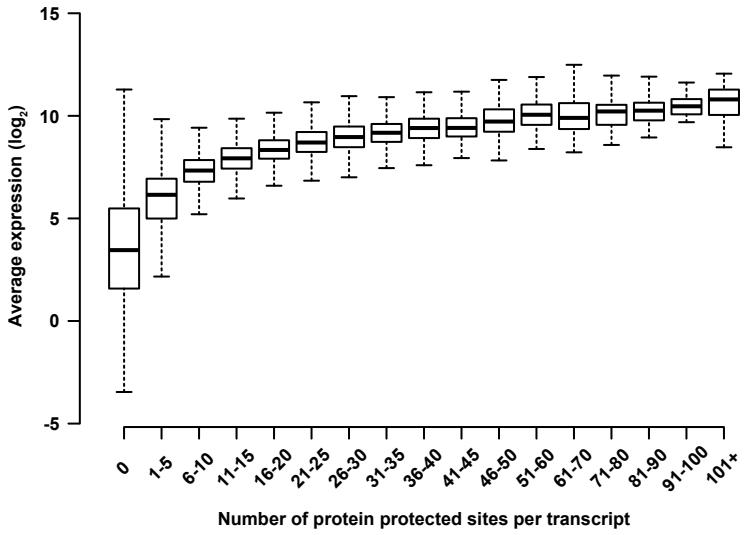


Figure 2.4 Average expression (y-axis) of human mRNAs separated by total number of PPSs identified in their sequence (x-axis) for formaldehyde cross-linking identified PPSs.

In general, we found that formaldehyde cross-linking revealed the highest number of PPSs, whereas UV and no cross-linking yielded many fewer sites (Table 2.1). This is not surprising, given that formaldehyde both has longer range than UV and also can stabilize more transient and indirect interactions. Thus, the use of formaldehyde cross-linking gives a more comprehensive view of RNA-protein interaction sites, while the use of UV likely increases the specificity of PPSs to more tightly associated RBP-bound targets. We also observed that ssRNase treatment yielded twice as many unique PPSs as compared to dsRNase digestion (Table 2.1). There are several explanations for this, none of which are mutually exclusive. For example, the ssRNase may have higher activity in the reaction conditions used in our experiments, the dsRNase may have lower accessibility to protein-bound dsRNA regions, or human RBPs may prefer non-structured regions within target RNAs for interaction. Together, these results show that the choice of cross-linking reagent or RNase can have a profound effect on RNA-protein interaction site identification and that these effects likely apply to the other technologies that address this same experimental question [127, 128].

2.2.2 PIP-seq is a reproducible approach known and novel RBP interaction sites

To assess the reproducibility of PIP-seq, we first determined the correlation of sequencing read abundance between biological replicates of footprinting and RNase digestion control libraries (Figure 2.5). Using a sliding window approach, we observed high correlation in read counts between individual replicates of formaldehyde cross-linked, ssRNase-treated footprinting and RNase digestion control libraries (Pearson correlation r = 0.88 and 0.84, respectively) (Figure 2.5A). Similar results were also found for the dsRNase treated libraries (Pearson correlation r = 0.84 and 0.76, footprinting and RNase digestion control, respectively) (Figure 2.5B). This high reproducibility of PIP-seq libraries was also observed between replicates of the UV cross-linked libraries (data not shown). Together, these data indicate that PIP-seq experiments and controls are reproducible across replicates using various RNases and cross-linkers.

We next investigated the reproducibility of exact PPS identification between paired biological replicates. With formaldehyde cross-linking, we observed 68% and 42% (for ssRNase and dsRNase, respectively) overlap between PPSs identified in two replicates (Figure 2.6A). Similarly, 73% and 64% (ssRNase and dsRNase, respectively) of the PPSs identified by UV cross-linking were replicated in a second, larger data set (Figure 2.6B). This degree of overlap between PPSs is relatively high when compared to the more modest reproducibility of the identified RBP binding sites in CLIP-seq and PAR-CLIP experiments [120]. In total, these results indicate that our novel approach is a reproducible means of identifying the protein-bound component of the eukaryotic transcriptome.

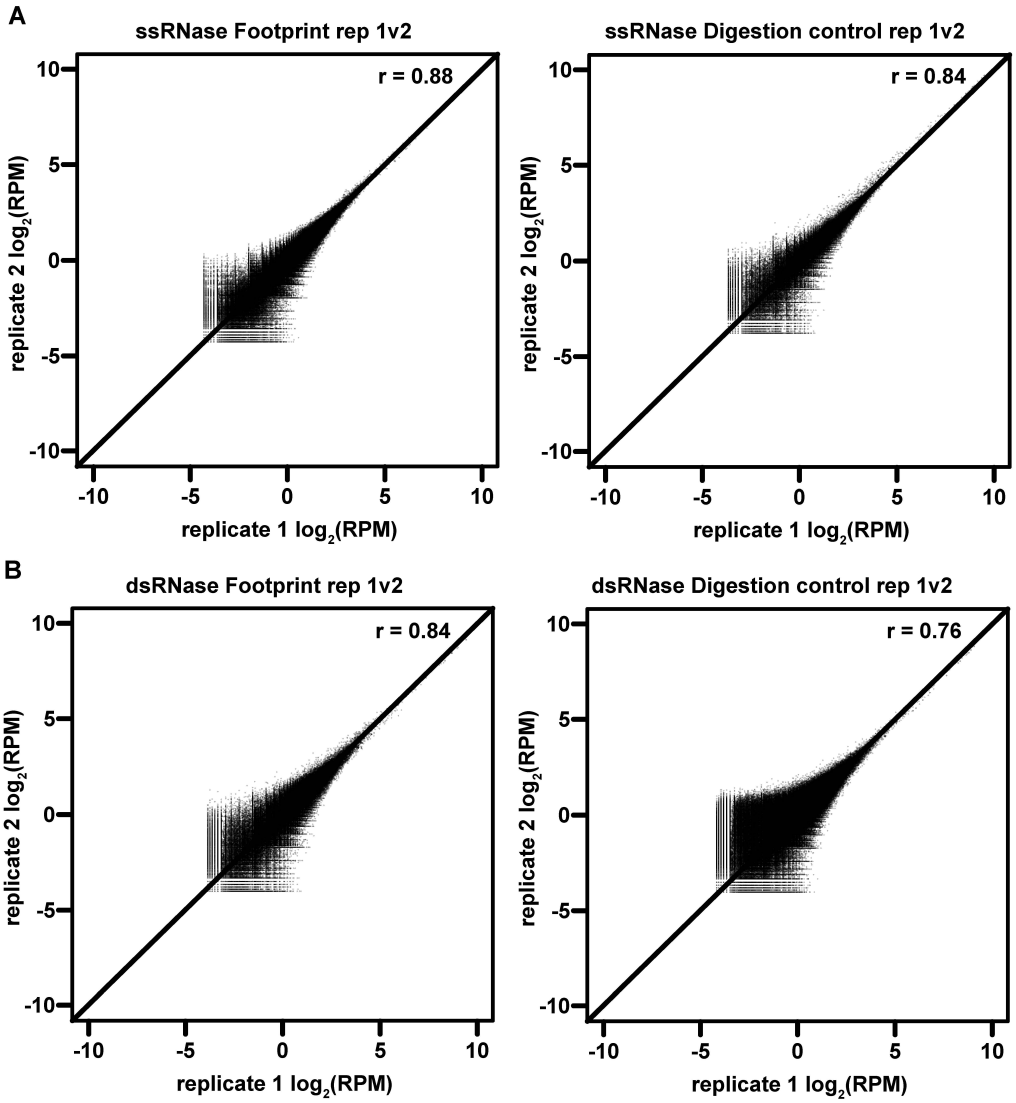


Figure 2.5 Correlation of PIP-seq replicates. A-B) Correlation in read counts between two formaldehyde cross-linked (A) ssRNase-treated PIP-seq replicates (footprinting sample on left, RNase digestion control on right). (B) As in (A), but for formaldehyde cross-linked dsRNase-treated replicates.

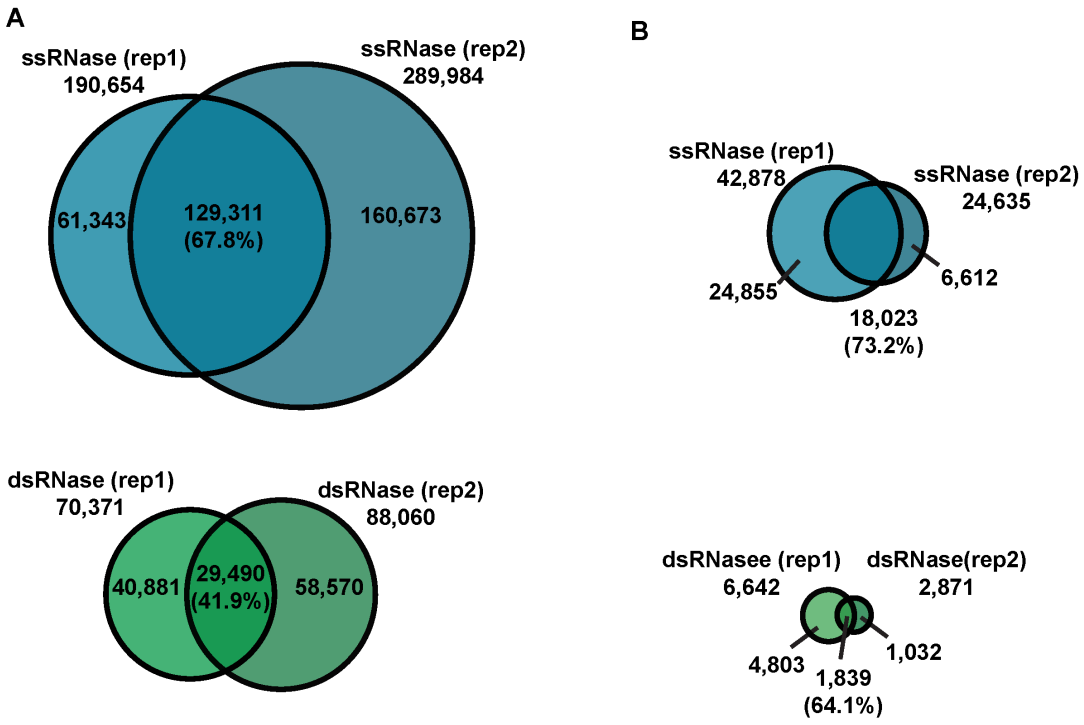


Figure 2.6 Overlap in PPS calls between formaldehyde (A) and UV-cross-linked (B) ssRNase-treated (top, blue), and formaldehyde cross-linked dsRNase-treated (bottom, green) PIP-seq replicates.

We also interrogated the relationship between PPSs identified by different RNases. We compared the use of RNaseONE, which preferentially cleaves single-stranded RNA, to RNaseV1, which preferentially cleaves paired bases (Figure 2.7). We found high overlap between formaldehyde PPSs (72%) identified by each RNase, as compared to UV (32%) or non-cross-linked (37%) PPSs. This is unsurprising, given the larger number (Table 2.1) of formaldehyde identified PPSs as compared to UV or non-cross-linked experiments. In total, these results revealed that both RNases uncovered a set of overlapping and unique PPS sequences, demonstrating that the use of an ss- and dsRNase is needed for comprehensive identification of RNA-protein interaction sites in eukaryotic transcriptomes.

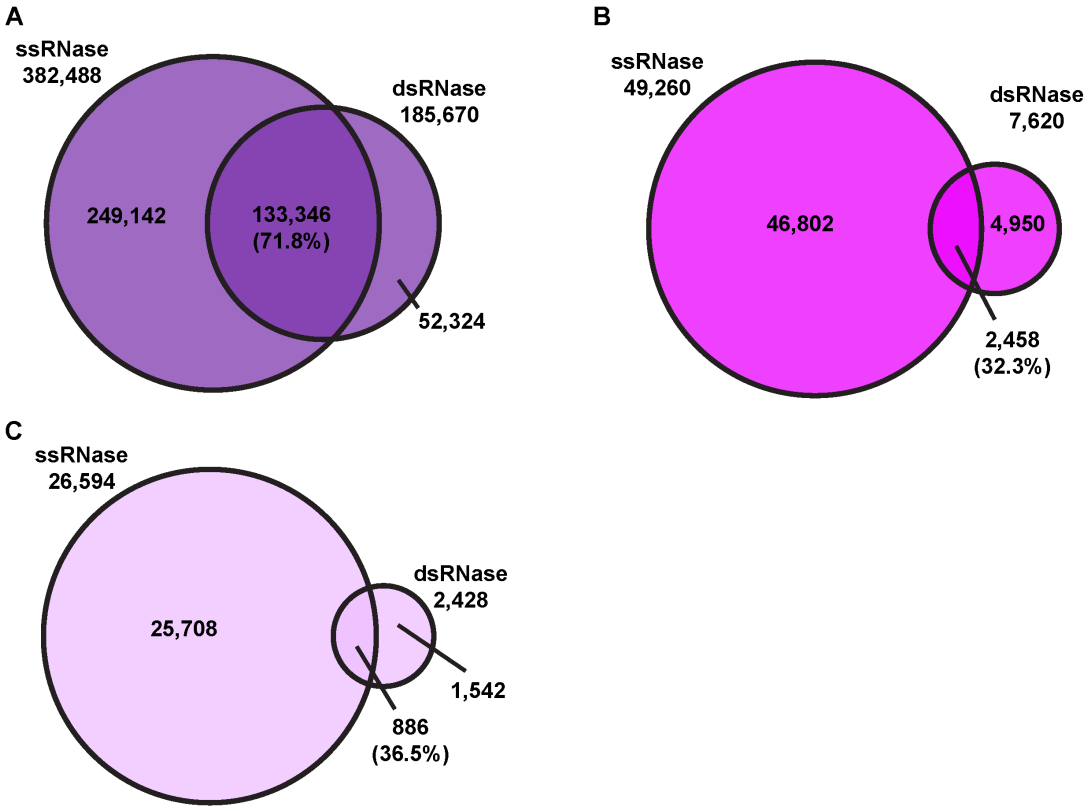


Figure 2.7 Overlap in PPS calls between formaldehyde (A), UV- (B) and non-cross-linked (C) ssRNase and dsRNase treated PIP-seq samples.

To validate that PIP-seq identifies bona fide RNA-protein interaction sites, we overlapped PPSs with known RBP binding sites from HeLa and HEK293T cells [89, 103, 116-127], and found that a significant number (all p-values < 2.2e-16) of the PPSs coincided with numerous RBPs previously tested by single protein immunoprecipitation approaches (e.g. HITS-CLIP, PAR-CLIP, etc.) as compared to an expressed transcriptome background (see Materials and Methods for more details) (Figure 2.8). This is noteworthy given our analysis of PPSs in HeLa cells, whereas the majority of the CLIP-seq and PAR-CLIP datasets were generated using HEK293T cells.

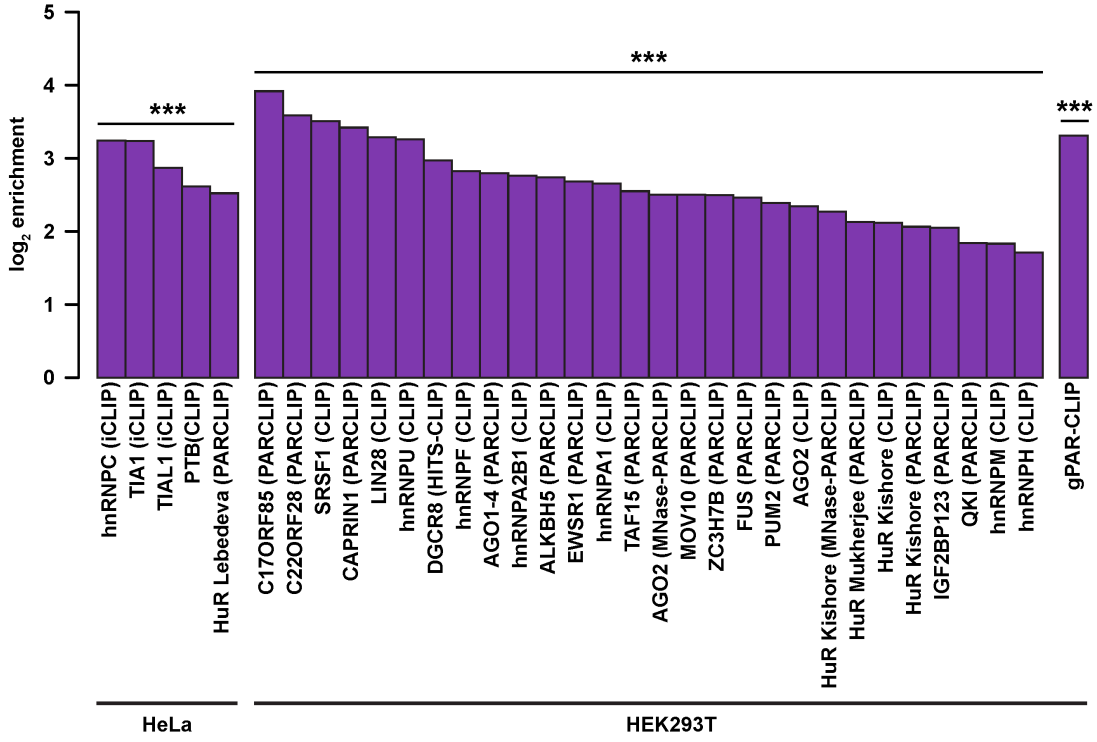


Figure 2.8 Overlap between PPSs identified from formaldehyde cross-linking treated PIP-seq samples and various CLIP datasets. Values are shown as $\log_2$ enrichment over shuffled background distributions. *** denotes p-value < 2.2e-16, Chi-squared test.

We also compared our data with previously published global PAR-CLIP (gPAR-CLIP) data from HEK293T cells [127], in which protein-binding sites were identified on the basis of T>C transversions (Figures 2.8 and 2.9). We observed a significant (p-value < 2.2e-16) enrichment of the previously identified transversion events within our identified PPSs relative to the expressed transcriptome background, suggesting that at least some fraction of binding events are cell type independent (~38% overlap between HeLa and HEK293T, Figures 2.8 and 2.9). Furthermore, we analyzed the number of T>C transversions per PPS and found that on average 6.3 T>C transversions were observed per PPS for the formaldehyde cross-linked PPSs (Figure 2.10). These data revealed that there are often numerous gPAR-CLIP T>C transversions per RNA-protein binding event identified by PIP-seq, and suggest that many of our identified PPSs

represent sites of multi-RBD and/or multi-RBP interactions. Additionally, our findings demonstrate that PIP-seq can identify the full footprint of RBP-RNA interaction sites, underscoring its utility in studying these events.

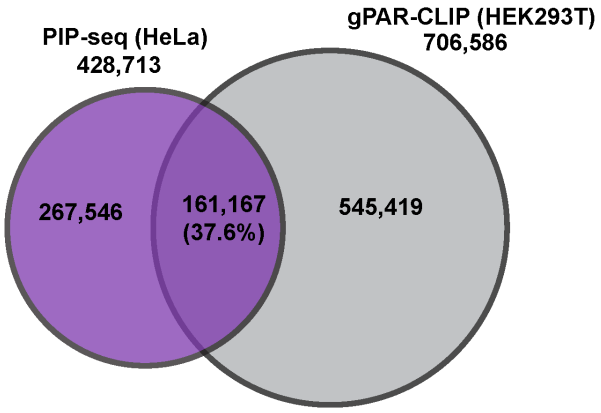


Figure 2.9 Overlap between cross-linked PPSs from HeLa cells and 40 nt T>C transversion event-containing loci from the gPAR-CLIP dataset generated from HEK293T cells (T>C transversion events less than 40bp apart were merged to generate a data set comparable to PPSs).

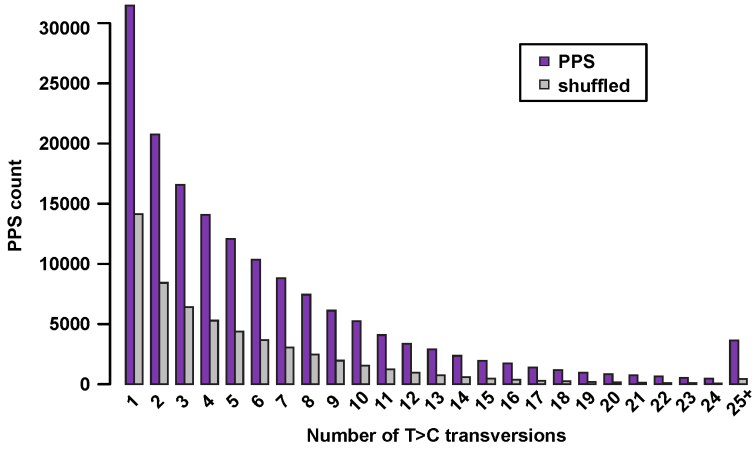


Figure 2.10 Number of T>C transversion events per PPS identified by formaldeyde cross-linking (purple) versus shuffled regions (gray). Values for the number of events per shuffled region are the average from ten random shuffles.

It is also worth noting that PIP-seq identified a total of 428,713 ~40 nt protein protected regions, while gPAR-CLIP yielded 706,586 loci of similar length (Figure 2.10). There are multiple explanations for this discrepancy. For instance, PIP-seq involves the use of a background control library (RNase digestion control (Figure 2.1)) whereas gPAR-CLIP does not. This control is likely important for distinguishing between noise and true protein binding events, and may account for the identification of fewer sites by PIP-seq. Alternatively, PIP-seq may be less sensitive due to the lack of a stringent RNA-protein purification step. In total, our results indicate that PIP-seq captures a significant population of human RNA-protein interaction regions in a single experiment, further validating its reliability and robustness.

2.2.3 PIP-seq reveals an in-depth view of the protein-bound transcriptome

Two outstanding questions in the field of RNA biology are the extent and patterning of RBP binding across genic regions. We set out to address these questions using PIP-seq data from the various cross-linkers and RNases. We first determined the size distribution of PPSs identified by each RNase and cross-linker (Figure 2.11). We found that the median PPS sizes for formaldehyde cross-linked ss- and dsRNase treatments were ~40 and ~35 nt, respectively. Importantly, this variation in size between the two RNases was consistent across cross-linkers (Figure 2.11), suggesting that ssRNase treatment reveals larger protein footprints and/or longer stretches of RBP interactions across RNA regions.

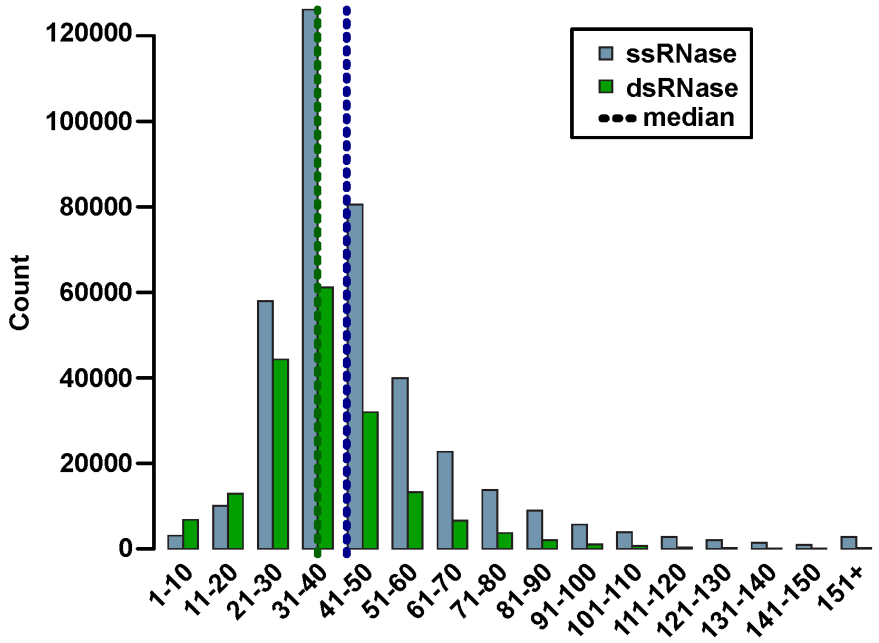


Figure 2.11 Distribution of ssRNase-treated (light blue bars) and dsRNase-treated (green bars) PPS sizes from formaldehyde treated samples. Dashed lines represent mean PPS sizes (ssRNase, blue line and dsRNase, green line).

To assess the genomic distribution of protein binding events, we calculated the enrichment of PPSs in specific regions of the human transcriptome (e.g. coding sequence [CDS], 5'UTR, 3'UTR, intron, etc.) relative to their expression levels in the RNase digestion control sample (Figure 2.12). This analysis revealed a consistent enrichment between RNases and cross-linkers for protein-binding in the 3'UTR, proximal (< 500 nt from a splice site) introns, as well as within the CDS (Figure 2.12). These results are unsurprising given the role of these regions in post-transcriptional regulation and translation. We also found that distal (> 500 nt from a splice site) intronic regions were enriched for protein binding in the formaldehyde treated samples only (Figure 2.12), suggesting a high level of transient, weak, and/or non-specific RNA-binding activity occurs in these non-coding areas. Our results support the idea that the large interior regions of introns may serve as sinks for RBPs in human cells [121].

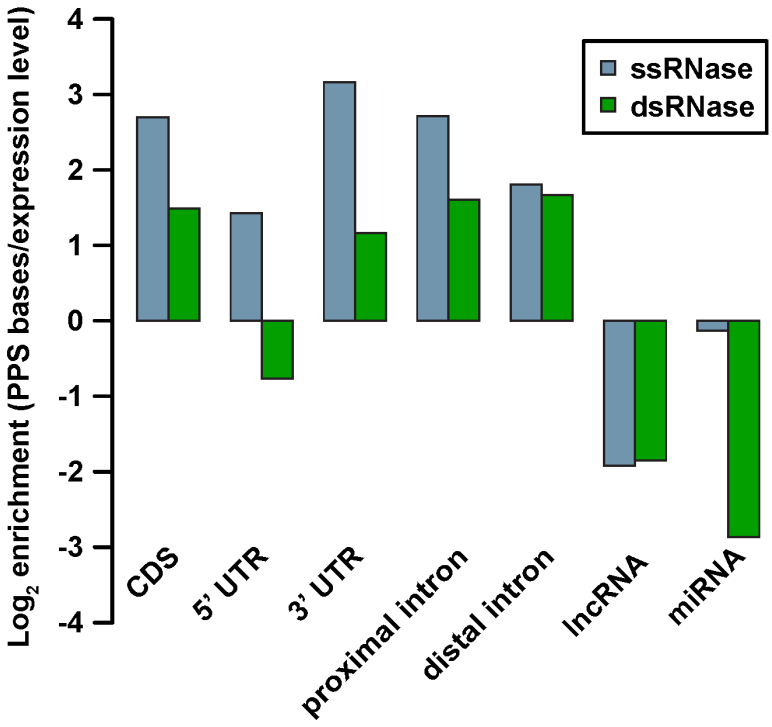


Figure 2.12 Genomic distribution of PPS density for formaldehyde crosslinked (C) samples, measured as PPS base coverage normalized to RNase digestion control read counts per genomic region. Proximal intron refers to 500 nt at the 5' and 3' ends of introns.

In contrast to protein-coding mRNAs, we found that long non-coding RNAs (lncRNAs) were consistently depleted for protein-binding (Figure 2.12). Therefore, we closely examined protein binding to the 100 most highly expressed lncRNAs compared to expression-matched mRNA 3' UTRs in the three different cross-linking conditions. These analyses revealed that the fraction of identified lncRNA and 3'UTR base pairs bound by proteins was similar for the formaldehyde cross-linking experiments using both RNases. Conversely, for UV and no cross-linking, lncRNAs demonstrated a significant depletion in protein binding compared to the expression-matched mRNA 3' UTRs (Figure 2.13). This depletion was consistent for both RNases, suggesting that this finding is not a consequence of structural differences between mRNAs and lncRNAs. In total, these results support the hypothesis that lncRNAs are more

weakly and/or transiently bound by interacting proteins as compared to protein-coding mRNAs, which may be a distinguishing feature of these two types of eukaryotic RNAs.

Given the fundamental role of RBP-RNA interactions in the regulation of eukaryotic gene expression, we hypothesized that many of the identified PPSs would be evolutionarily conserved within vertebrates. To test this, we compared SiPhy-π conservation scores for PPSs versus same-sized neighboring regions (Figure 2.14). Using this approach, we found that PPS sequences were significantly (p-value < 2.2e-16) more evolutionarily conserved than flanking regions (Figure 2.14A-C). Importantly, this was true for PPS sequences in both exonic and intronic portions of human mRNAs, but not for lncRNAs (Figures 2.14D and G), and was consistent for PPSs identified with every cross-linking approach (Figures 2.14E-F and H-I). These results support the notion that the ability to interact with RBPs is functionally important to mRNA sequences, and that this trait has undergone selection during vertebrate evolution. Furthermore, the lack of conservation of PPSs within lncRNAs is consistent with their low conservation rates across vertebrate species.

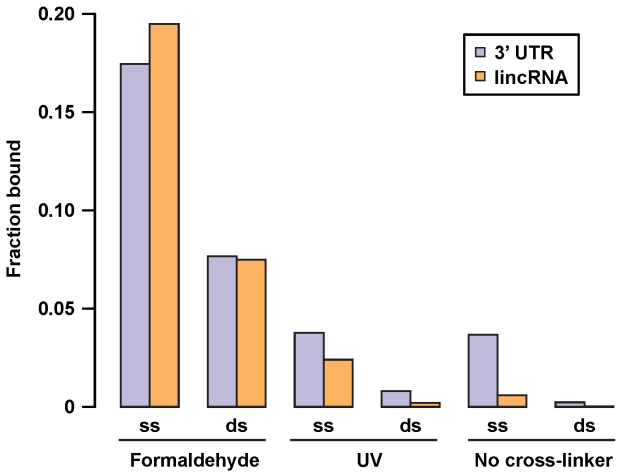


Figure 2.13 Fraction of base pairs covered by PPSs in 100 most highly expressed lncRNAs (orange bars) and expression-matched control mRNA 3'UTRs (purple bars) for PIP-seq libraries made with ssRNase (ss) or dsRNase (ds) under the three different cross-linking conditions (as specified).

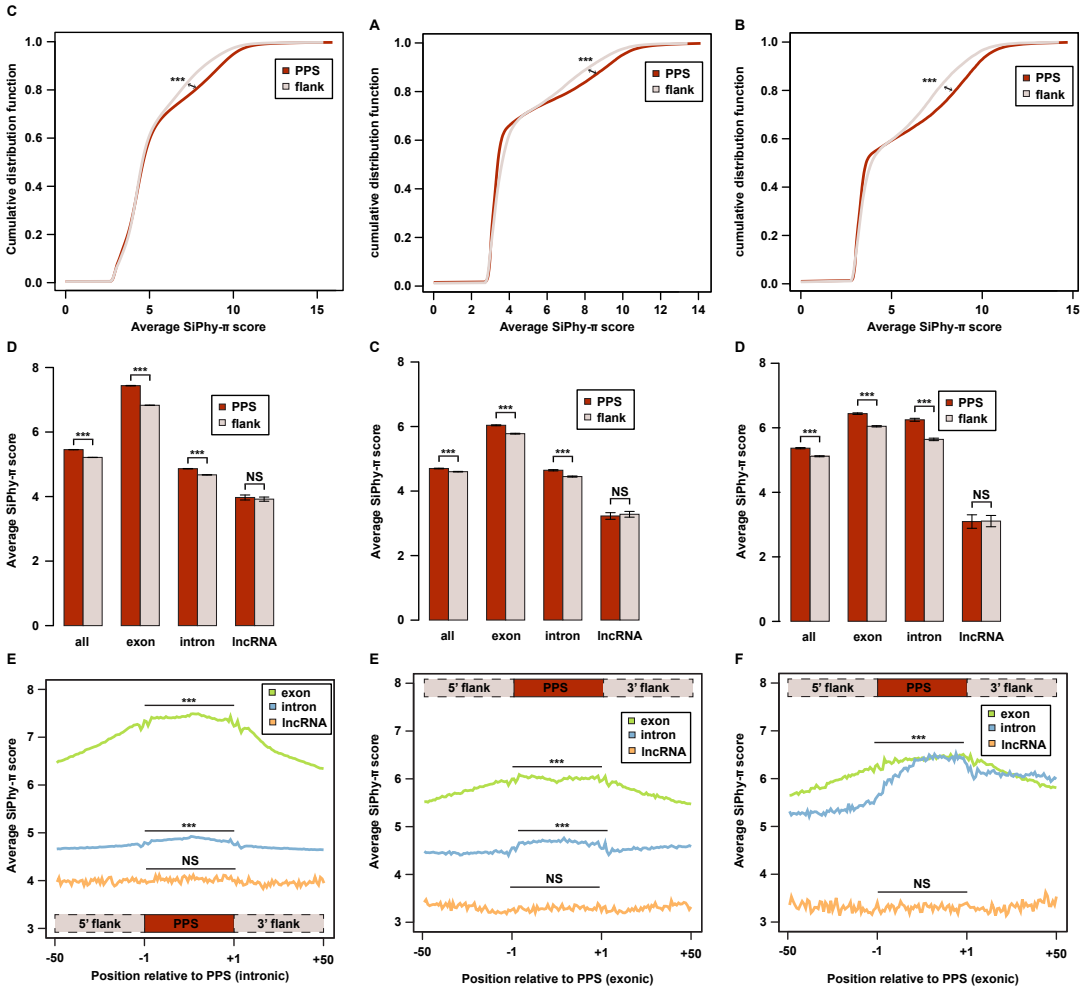


Figure 2.14 PPSs are highly conserved. (A-C) Cumulative distribution of average SiPhy-π scores in formaldehyde (A), UV- (B) and non-cross-linked (C) identified PPSs (red line) versus similarly-sized flanking sequences (gray line). (D-F) Comparison of average SiPhy-π scores between formaldehyde (D), UV- (E) and non-cross-linked (F) identified PPSs (red bars) and flanking sequences (gray bars) for various genomic regions. (G-I) Average SiPhy-π score profiles across the first and last 25 nt of formaldehyde (G), UV- (H) and non-cross-linked (I) identified PPSs as well as 50 nt upstream and downstream of exonic (green line), intronic (blue line), and lncRNA (orange line) PPSs. *** denotes p-value < 2.2e16, NS = not significant, Chi-squared test.

2.2.4 RBP binding densities across unprocessed and mature mRNAs

Given the importance of RBP binding within different regions of mRNAs, we decided to determine the density of protein-binding sites within specific regions of protein-coding transcripts (Figure 2.15). To do this, we first identified PPSs within each annotated CDS, 5' UTR, 3'UTR, and intronic region and calculated the relative distribution of binding sites across these regions (Figures 2.15A-C). We corrected for average length of each region to obtain a global view of relative binding between regions. We also calculated PPS coverage on a per nucleotide basis for specific sub-regions of protein-coding mRNAs (Figures 2.15D-I).

Applying this approach to PPSs identified with formaldehyde cross-linking, we observed similarly high levels of binding within the entirety of the CDS and 3' UTR of protein-coding transcripts with an enrichment for binding events occurring at and near the start and stop codons (Figures 2.15A and D). This enrichment was particularly evident when interrogating the PPS density over the start and stop codons on a per nucleotide basis (Figure 2.15D). Similar enrichments leading to the start of the CDS were identified when defining PPS densities in the 5' UTR. We also found that the overall protein binding density was lower in the 5' UTR when compared to the CDS and 3' UTR (Figures 2.15A). The observed enrichment of PPSs at the CDS start and stop codon regions likely reflects ribosome binding, as was previously observed by others [127, 128].

Overall similar patterns of RBP binding were also observed for the UV and no cross-linking experiments (Figures 2.15B-C). The two exceptions were that UV and non-cross-linked RBP binding density across the 3' UTR peaked near the middle of this region (Figures 2.15B-C), and the interaction profile directly over the start codon displayed a minor depletion in protein binding in these experiments (Figures 2.15E-F). These results likely reflect the differential cross-linking specificities of formaldehyde and UV, and support the use of multiple cross-linkers in the comprehensive identification of RBP binding sites.

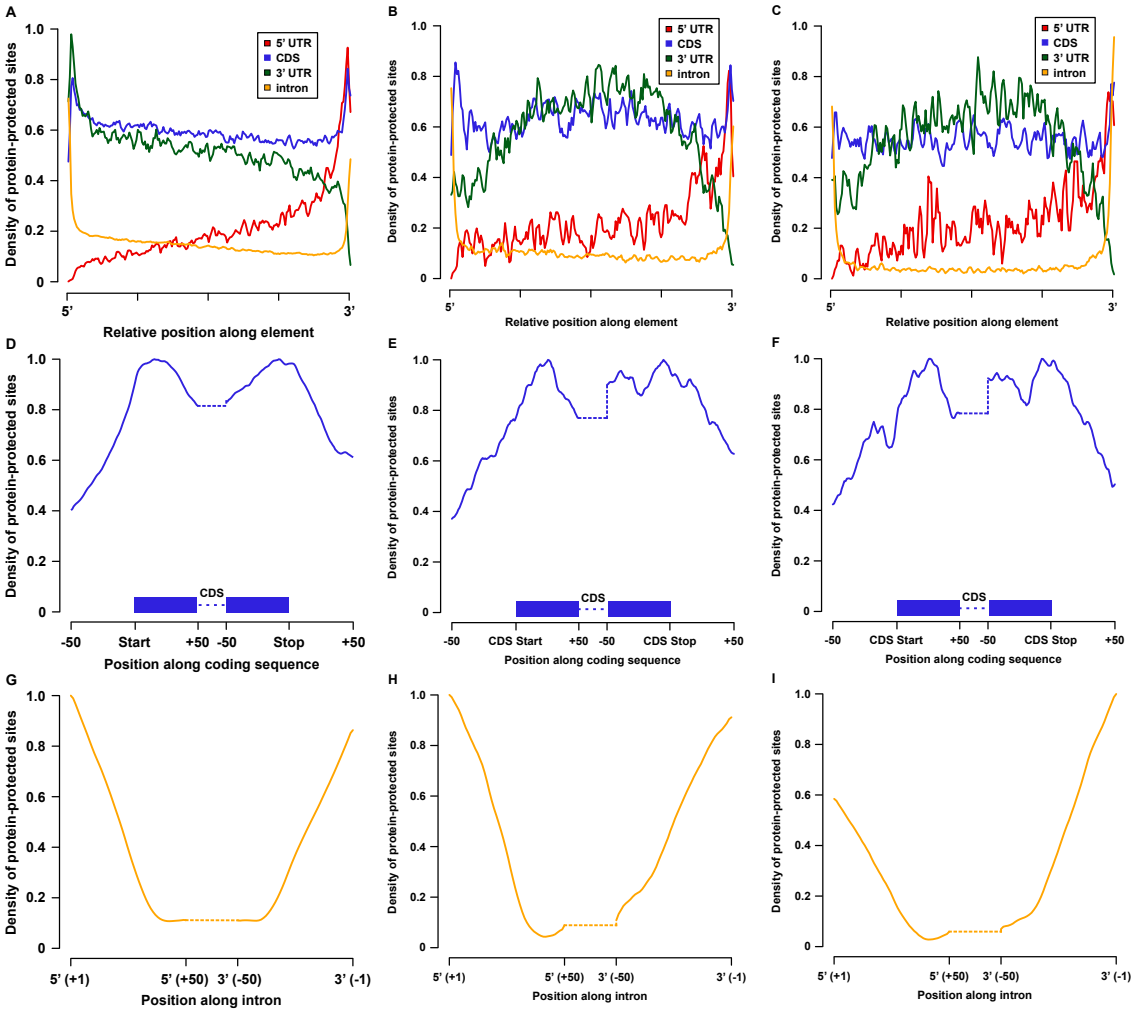


Figure 2.15 PPS coverage across mRNAs. (A-C) Average PPS density for formaldehyde (A), UV- (B) and non-cross-linking (C) experiments across 100 equally spaced bins in various genic regions. Values are normalized separately for each genic region (e.g. intron). (D-F) Average PPS density for formaldehyde (D), UV- (E) and non-cross-linking (F) experiments within 50 nt of CDS ends. (G-I) Average PPS density for formaldehyde (G), UV- (H) and non-cross-linking (I) experiments within the first and last 50 nt of introns. Dotted lines in (D-I) represent the remaining (unanalyzed) length of each element.

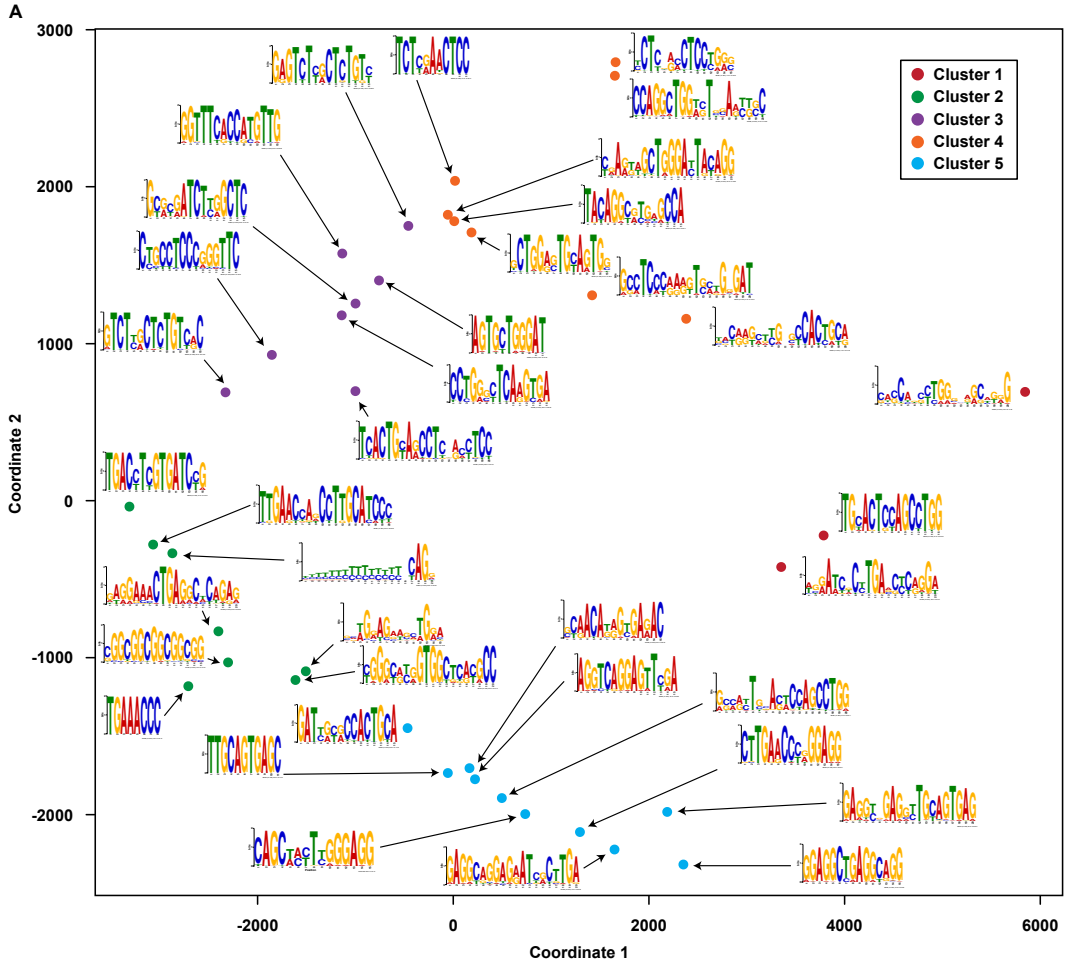
Given the ability of PIP-seq to capture unprocessed RNAs, we also investigated RBP binding density across introns. Unsurprisingly, we observed most binding events proximal to the 5' and 3' splice sites (Figures 2.15A-C). This was consistent across cross-linkers and is likely due to extensive association with lariat formation machinery proximal to the splice sites. At single

base resolution, we located the beginning of this enrichment starting 40 nt away from each splice site, consistent with the binding location of RNA splicing factors (Figures 2.15G-I). In total, our results indicate that PIP-seq gives a comprehensive view of RNA-protein interaction site densities in all portions of mature as well as unprocessed mRNAs, especially when multiple cross-linking agents are employed.

2.2.5 PIP-seq provides evidence for the post-transcriptional operon hypothesis

Given that PPSs correspond to protein-bound RNA sequences (Figure 2.8), we sought to gain insights into the sequence elements that are enriched within RNA-protein interaction sites in the HeLa transcriptome. To do this, we employed the MEME (Mulitple EM for Motif Elicitation) algorithm [135] on PPSs partitioned by specific region (e.g. 5' UTR, 3' UTR, CDS, and intron). Because we could not rule out ribosome binding at start and stop codons, we additionally removed the first and last exon of each CDS. Using this approach, we identified previously known binding motifs including sequences similar to the LIN28 binding motif [124] and U-rich sequences (accessible at gregorylab.bio.upenn.edu/PIPseq). We also identified numerous putative RBP binding motifs, some of which are particularly interesting because they are long (~20 nt) and contain multiple strong consensus sequences flanked by weaker ones (3'UTR motifs 4 and 31 and intron motifs 1 and 13) (accessible at gregorylab.bio.upenn.edu/PIPseq). These motifs may correspond to binding by multiple RNA-binding domains (e.g. RRM) of a single protein or by a complex of multiple RBPs. Importantly, motifs with this signature have not been previously reported in CLIP-seq and PAR-CLIP data. In addition, we identified at least one sequence that displayed a high degree of self-complementarity (3' UTR motif 1). This is surprising, given that MEME does not use RNA secondary structure as a search feature while identifying motifs from a set of given sequences. These findings underscore the utility of PIP-seq and its use of multiple structure-specific nucleases to uncover hidden features of the protein-interacting transcriptome.

A



B

	GO Term	p-value
Cluster 1	positive regulation of multicellular organismal process	9.2e-3
	immune response	1.0e-2
Cluster 3	regulation of transcription	5.4e-3
	regulation of RNA metabolic process	1.0e-2
Cluster 4	positive regulation of caspase activity	1.3e-3
	positive regulation of peptidase activity	1.3e-3
Cluster 5	apoptosis	1.0e-2
	programmed cell death	1.0e-2

Figure 2.16 PPS analysis reveals evidence for post-transcriptional operons. (A) MDS analysis of RBP-bound motif co-occurrence in human mRNAs. The motifs used for this study were identified by a MEME-based analysis of PPS sequences. Sequences for all of the motifs used in this analysis can be found in Additional File 10. Colors indicate cluster membership as defined by k-means clustering (k = 5). (B) The most significantly enriched biological processes (and corresponding p-value) for target transcripts, where the specified clusters of motifs identified in (A) are co-bound.

Although RNAs are thought to be bound and regulated by multiple RBPs, very little is known about these interactions and the relationships between specific RBPs and their corresponding sequence motifs. To address this, we interrogated the interactions between putative RBP binding motifs (Figure 2.16A) discovered by our PIP-seq approach, since these are protein-bound sequences in HeLa cells. To do this, we first identified all instances of each motif within the global set of identified PPSs on target RNAs using FIMO [136]. We collapsed motifs with similar sequences and excluded those that were long (~20 nts) and non-degenerate because these likely represent repetitive sequences instead of true binding motifs. We then quantified the co-binding of the remaining motifs (~40) within all protein-coding mRNAs by counting the number of transcripts on which each pair of motifs was jointly found within PPSs. We then used k-means clustering of the resultant weighted adjacency matrix and identified 5 clusters of motifs that interact on highly similar sets of target mRNAs (Figure 2.16A). These findings indicate that many mRNAs contain numerous RBP interacting motifs within their sequences and that coordinated binding of RBPs to specific target transcripts may represent a general phenomenon of cellular RNA-protein interactions, as was previously proposed by the post-transcriptional operon hypothesis [39, 113].

We also used DAVID [137] to interrogate over-represented biological processes for RNAs that contained binding events for each motif from the five clusters identified in the k-means analysis (Figure 2.16A, Clusters 1, 3 – 5). It is of note that the motifs in Cluster 2 did not co-occur in a large enough group of bound transcripts to allow meaningful Gene Ontology (GO) analysis. We found that the most highly over-represented functional terms for the RNAs that contained these co-occurring sequence motifs in HeLa Clusters 1, 3 – 5 were related to distinct processes, including developmental processes and immunity (Cluster 1), caspase activity and apoptosis (Clusters 4 and 5, respectively), as well as regulation of transcription and RNA metabolic processes (Cluster 3) (Figure 2.16B). These results suggest that there are distinct groups of RBP recognition motifs that are involved in the post-transcriptional regulation of various collections of mRNAs encoding functionally related proteins.

2.2.6 Disease-linked SNPs correlate with protein-bound RNA sequences

A growing set of evidence suggests that multiple RNA-level mechanisms, some of which depend upon RNA-protein interactions, are the means by which particular single nucleotide polymorphisms (SNPs) in mRNAs effect human disease phenotypes [138-141]. In support of this concept, we found PPSs to be enriched in disease-associated SNPs from dbSNP build 137 and the NHGRI GWAS Catalog (Figure 2.17A). Furthermore, the ratio of synonymous to non-synonymous SNPs was also significantly higher within PPSs compared with the expressed transcriptome background (Figure 2.17B, p-value = 9.8e-04), lending further support to the notion that disruption of RNA-protein interactions underlies the disease mechanism of the polymorphisms in question.

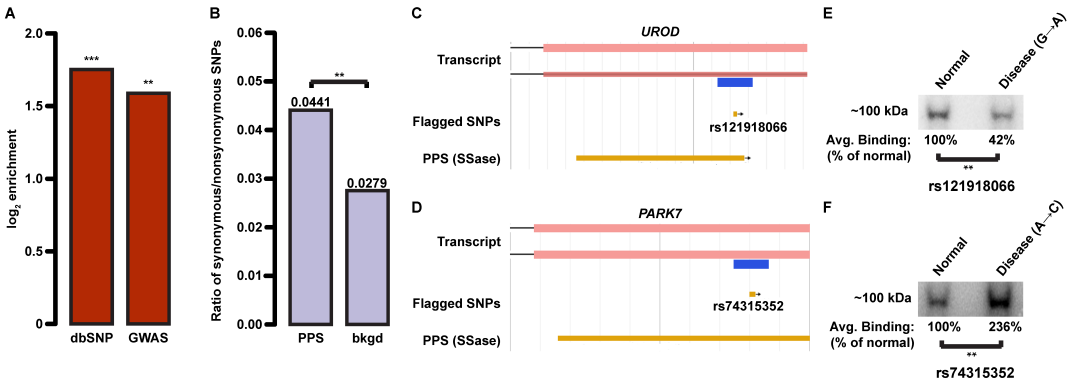


Figure 2.17 PPSs are enriched within disease-associated SNPs. (A) Enrichment of disease-associated SNPs from dbSNP build 137 and the NHGRI GWAS Catalog in PPSs versus background. *** denotes p-value → 0 and ** denotes p-value < 0.001, Chi-squared test. (B) Ratio of synonymous to non-synonymous SNPs in PPSs versus background. ** denotes p-value < 0.001, Chi-squared test. (C – D) Two examples of disease-related SNPs found in *UROD* (C) and *PARK7* (D) that overlap with PPSs identified by PIP-seq in HeLa cells using ssRNase treatment (SSase). The *UROD* and *PARK7* SNPs (as indicated in Flagged SNPs track) are used in the analyses in E – F, respectively. A blue line below the transcript model denotes the regions used for the analyses in E – F. (E – F) UV cross-linking analysis of normal compared to disease-related SNPs using probes with only the specific base pair substitution specified in parentheses next to disease label and protein lysates from HeLa cells. The rs121918066 (E) and rs74315352 (F) SNPs associated with Porphyria Cutanea Tarda and early-onset Parkinson's disease, respectively, were used in this analysis. Representative images for three replicate experiments. ** denotes p-value < 0.001. p-values were calculated by a one-tailed t-test.

To verify that disease-related human SNPs could affect RBP-RNA interactions, we used UV cross-linking analyses with 38 nt RNA probes containing either the normal or disease-associated variant at their center. For these analyses, we focused on two specific SNPs that are associated with Porphyria Cutanea Tarda and early-onset Parkinson's disease (rs121918066 and rs74315352, respectively). We found that both disease-associated SNPs tested had significant effects on specific RBP-RNA interactions (p-values < 0.001) (Figures 2.17C – D). In fact, we found that rs121918066 disrupted while rs74315352 enhanced specific interactions with an RBP complex. These findings revealed that disease-associated SNPs that reside within RBP binding sites can affect the interaction between proteins and their target RNAs. In total, these results suggest that modulation of RBP interactions may be a significant RNA-level disease mechanism in humans.

2.3 CONCLUSIONS

In general, the global architecture of RNA-protein interactions within the population of both unprocessed and mature RNA molecules is still poorly characterized [20, 114, 115]. Here, we described a novel RNase-mediated protein footprint sequencing approach (PIP-seq) that allows global identification of RNA-protein interactions for numerous RBPs in the human transcriptome with a single experiment (Figure 2.1). Our approach is similar to other recently published methodologies [127], but in addition to polyA-containing, mature mRNAs we also provide a view of RNA-protein interaction sites in unprocessed mRNAs (i.e. introns). Additionally, our approach is widely applicable to all samples and organisms since it is not dependent on the incorporation of non-natural nucleotides or UV cross-linking.

Analysis of the PPSs uncovered by our approach allowed us to identify significant levels of known and novel RNA-protein interaction sites and sequence motifs. By comparing across cross-linkers and RNases, we demonstrated that each uncovers specific subsets of protein-bound sequences and support the use of multiple reagents when obtaining a comprehensive analysis of the protein-bound transcriptome in eukaryotic organisms.

Using the RNA sequences identified as being protein-bound in the HeLa cell transcriptome by PIP-seq, we uncovered a large set of putative RBP-binding motifs. Based on their size and sequence characteristics, it is likely that many of these motifs correspond to binding sites for RBPs that interact with target RNAs through multiple RNA-binding domains or complexes of multiple RBPs. We used these identified RBP-bound motifs to investigate the interaction between RBPs within target mRNAs and offer insights into mRNP organization in the human transcriptome. This study is one of the first to comprehensively examine the co-binding by RBPs with specific target mRNAs. In fact, our findings provide an important resource for investigation into the idea that groups of RBPs bind to collections of mRNAs encoding proteins functioning in specific biological processes. These sequences can be used for identification of the interacting proteins so that their effects on post-transcriptional regulation can be further studied.

Finally, we observed a significant overlap of PPSs with disease-linked SNPs obtained from two different sources (dbSNP build 137 and NHGRI GWAS Catalog [142]), and validated these results using UV cross-linking experiments that demonstrated disease-linked SNPs could both disrupt or enhance RBP-RNA interactions. Thus, determining the molecular details behind each disease-associated SNP that affects an RNA-RBP interaction will be an important future research endeavor. It is also worth noting that our findings point to the intriguing possibility that PIP-seq could be used in conjunction with genome-wide association studies to screen for synonymous mutations that may be causal via altering of any number of RNA-protein interactions in affected tissues. Such a tool would be extremely valuable in mechanistic, pharmacogenomic, and therapeutic studies of disease-associated polymorphisms. In summary, we present a powerful method that will be important for future studies of RNA-protein interaction site dynamics in multiple eukaryotic organisms and in important biological contexts.

2.4 MATERIALS AND METHODS

Cell lines

For these experiments, HeLa cells were seeded in 15 cm Corning Standard tissue-culture treated culture dishes (Sigma, St. Louis, MO), grown to 90% confluence (~18 million cells) in DMEM media (Life Technologies, San Diego, CA) supplemented with L-glutamine, 4.5 g/L D-Glucose, 10% FBS serum (Atlanta Biologics, Atlanta, GA), and Pen/Strep (Fisher Scientific, Waltham, MA).

Cross-linking experiments

For formaldehyde cross-linking, 37% formaldehyde solution (Sigma, St. Louis, MO) was added drop-wise with mixing directly to cell culture dishes containing 90% confluent cells to a final concentration of 1% and incubated at room temperature for 10 minutes. Next, 1M glycine (Sigma, St. Louis, MO) was added to a final concentration of 125 mM and incubated for an additional 5 minutes with mixing. Then, cells were washed twice with ice cold PBS and collected. Finally, cells were pelleted and frozen until the PIP-seq digestions were performed. For UV cross-linking experiments, 90% confluent cells were washed twice with ice cold PBS and resuspended in 5 ml of PBS. Cell culture dishes were placed in a UV Stratalinker 2400 (Agilent Technologies, New Castle, DE) with the lid removed and irradiated with UV-C (254 nm) once with 400 mJ/cm^2. The cross-linked cells were collected by scraping, pelleted, and then frozen until used.

PIP-seq library preparation

To begin, we lysed the cell pellets 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 protease inhibitors/10 ml) and manual grinding (850 μl of RIP is used per 10 million cells). The resulting cell lysate was treated with

RNase-free DNase (Qiagen; Valencia, CA). Subsequently, these DNA-depleted lysates were split and treated with either 100 U/ml of a single-stranded RNase (ssRNase) (RNaseONE (Promega; Madison, WI)) with 200 µg/ml BSA in 1X RNaseONE buffer for 1 hour at room temperature, or 2.5 U/ml of a double-stranded RNase (dsRNase) (RNaseV1 (Ambion; Austin, TX)) in 1X RNA structure buffer for 1 hour at 37°C as previously described ([129, 130], see Figure 2.1 for a schematic description). 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 room temperature. It is worth noting for clarity that we had two cell lysates for these experiments: one treated with the ssRNase and the other with dsRNase. For formaldehyde cross-linking experiments, proteinase digestion was followed by a 2-hour incubation at 65°C to reverse the cross-links, whereas for UV cross-linking experiments, RNA was liberated from protein by retreating the lysates with 1% SDS and 1 mg/ml Proteinase K for 30 minutes.

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 in the absence of bound proteins. To do this, we performed the identical treatments as described above except that the cross-linked cellular 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 (see Figure 2.1 for schematic).

The digested RNA was then isolated using the Qiagen miRNeasy RNA isolation kit following the included protocol (Qiagen; Valencia, CA). Finally, the purified RNA was used as the substrate for strand-specific sequencing library preparation, as previously described [129, 130], with the exception that we also include DSN library normalization per manufacturer instructions (Illumina; San Diego, CA). Briefly, 100 ng of the final library was denatured at 95°C and then annealed for 5 hours at 68°C. 2 µl of DSN enzyme (1U/µl) was used to deplete re-annealed

duplexes. All of the RNase footprinting libraries (a total of 4 for each replicate (ss- and dsRNase treatments, footprint and RNase digestion controls)) were sequenced on an Illumina HiSeq2000 using the standard protocols for 50 base pair (bp), single read sequencing**.**

Read processing and alignment

PIP-seq reads were first trimmed to remove 3' sequencing adapters using cutadapt (version 1.0 with parameters -e 0.06 -O 6 -m 14). The resulting trimmed sequences were collapsed to unique reads and aligned to the human genome (hg19) using Tophat (version 2.0.9 with parameters --read-mismatches 2 --read-edit-dist 2 --max-multihits 10 --b2-very-sensitive --transcriptome-max-hits 10 --no-coverage-search --no-novel-juncs). PCR duplicates were collapsed to single reads for all subsequent analyses.

Identification of PPSs

PPSs were identified using a modified version of the CSAR software package [134]. 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 RNase digestion control libraries. PPSs were then called as described [134] with an FDR of 5%.

PPS saturation analysis

Mapped reads from chromosome 9 of formaldehyde cross-linked ssRNase treated PIP-seq replicate 1 libraries were randomly subsampled at 10%-90% by a custom perl script. CSAR was used to identify PPSs as described and total number of PPSs was plotted as a function of subsample size.

Validation by comparison with CLIP-seq, PAR-CLIP and gPAR-CLIP data

iCLIP, PAR-CLIP, and CLIP-seq datasets were compiled from sources as referenced and overlapped with PPSs. Significance of overlaps with PPSs was assessed using a Chi-squared test compared to an expressed transcriptome background. To compute a background distribution for the number of T>C transversions, we generated 10 random sets of genomic intervals with the same size distribution as PPSs. These random intervals were selected from a background of actively transcribed regions (defined using bgrSegmenter [143] with parameters threshold=10 maxGap=10 minRun=15).

Functional analysis of PPSs

Gene annotations were downloaded from the UCSC Genome Browser (RefSeq Genes, wgRna, rnaGene, lncRNA), and miRBase release 18 was used for microRNA annotations. PPS annotation was done 'greedily', such that all functional annotations that overlapped with a given PPS were counted equally. Conservation was assessed by computing average SiPhy-π log-odds [144] scores within PPSs and in equally-sized regions immediately upstream and downstream of each PPS.

Motif and co-occurrence analysis

MEME [135] was used to identify enriched RBP interaction motifs with parameters –dna –nmotifs 100 –evt 0.01 –maxsize 100000000. Motif co-occurrence was defined at the transcript level, and k-means clustering of the resultant weighted adjacency matrix was used to identify modules of co-occurring motifs. We set k=5 based on manual inspection of clusters on a multidimensional scaling (MDS) plot of the adjacency matrix. Gene Ontology (GO) analysis was performed using DAVID [137].

Analysis of SNPs and disease associations

Clinically associated SNPs (snp137Flagged) were downloaded from the UCSC Table Browser. We also downloaded the NHGRI GWAS Catalog [142] of disease-linked SNPs. Background distributions refer to the incidence of each dataset within the same genic regions as those of the PPSs in each analysis. Significance was assessed using a Chi-squared test.

UV Cross-linking analysis of disease-associated SNPs

We generated asymmetric oligonucleotide hybrids for *in vitro* transcription by annealing T7 sense DNA oligonucleotides (TAATACGACTCACTATAGGG) to antisense probe sequences fused to the antisense T7 (aT7) sequence (rs74315352 normal: CTTGTAAGAATCAGGCCGtCTTTTTCCACACGATTCTC(aT7), rs74315352 disease: CTTGTAAGAATCAGGCCGgCTTTTTCCACACGATTCTC(aT7), rs121918066 normal: CCCAGGTTGGCAATGTAGcGATGTGGTCCAAAGTCATC(aT7), rs121918066 disease: CCCAGGTTGGCAATGTAGtGATGTGGTCCAAAGTCATC(aT7)) (IDT; San Jose, CA). Each hybrid reaction was incubated at 95°C for five minutes and cooled to 25°C by step-wise increments of 1°C/minute.

In vitro transcription reactions were performed by adding 1 μg of the asymmetric oligonucleotide hybrids (see above) to a 25 μL transcription reaction comprising 1X T7 RNA Transcription buffer (NEB, Cambridge, MA), 36 μM UTP (for rs74315352) or 36 μM CTP (for rs121918066), 264 μM each of ATP, CTP and GTP (for rs74315352) or 264 μM each of ATP, UTP and GTP (for rs121918066), 0.04 mCi ^{32}P UTP (for rs74315352) or 0.04mCi ^{32}P CTP (for rs121918066), 10 nM DTT, 40 U RNaseOUT (Invitrogen; Carlsbad, CA), and 75 U of T7 RNA Polymerase. The reactions were incubated at 37°C for two hours. DNA was digested with 4 units of Turbo DNase (Invitrogen; Carlsbad, CA) at 37°C for 20 minutes. RNA probes were chloroform extracted and precipitated. The amount of labeled RNA probe was determined by 15% TBE-Urea

gel electrophoresis followed by phosphorimaging and densitometry. Normal and disease RNA probes were normalized to equal activities and used for subsequent analysis.

Equal concentrations of each RNA probe (~10% of total from *in vitro* transcription) were 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 1.5 μg/μl HeLa whole cell 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 protease inhbitors/10mL) 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)). To digest unbound RNA, each reaction was incubated with 20 U RNase T1 and 8 μg RNase A at 37°C for 20 minutes. 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 (Invitrogen; Carlsbad, CA) at 130V for 1.5 hrs. Phosphorimaging and densitometry were used to visualize and quantify protein-bound RNA, respectively.

Accession Numbers

All PIP-seq data from our analyses were deposited in GEO under the accession GSE49309. All of our data (i.e. files of all identified PPSs, complete lists of overrepresented motifs, GO analyses, etc.) can also be accessed at http://gregorylab.bio.upenn.edu/PIPseq/. Web browsers for visualization of all PPSs and our analyzed and raw sequencing data can be found at http://gregorylab.bio.upenn.edu/jbrowse/?data=data/HeLa_PIPseq for jbrowse, and at http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=pipseq&hgS_otherUserSessionName=PPS for the UCSC genome browser.

CHAPTER 3: PABPC1 BINDS TO GENOMICALLY ENCODED SEQUENCES WITHIN MAMMALLIAN MRNAS

This section refers to work from:

- Kini H*, Silverman IM*, Ji X, Gregory BD, Liebhaber SA. Cytoplasmic poly(A) binding protein-1 binds to genomically encoded sequences within mammalian mRNAs. *RNA*

Abstract:

The functions of the major mammalian cytoplasmic poly(A) binding protein, PABPC1, have been characterized predominantly in the context of its binding to the 3' poly(A) tails of mRNAs. These interactions play important roles in post-transcriptional gene regulation by enhancing translation and mRNA stability. Here, we performed transcriptome-wide CLIP-seq analysis to identify additional PABPC1 binding sites within genomically encoded mRNA sequences that may impact on gene regulation. From this analysis, we found that PABPC1 binds directly to the canonical polyadenylation signal in thousands of mRNAs in the mouse transcriptome. PABPC1 binding also maps to translation initiation and termination sites bracketing open reading frames, exemplified most dramatically in replication-dependent histone mRNAs. Additionally, a more restricted subset of PABPC1 interaction sites comprised A-rich sequences within the 5' UTRs of mRNAs, including *Pabpc1* mRNA itself. Functional analyses revealed that these PABPC1 interactions in the 5'UTR mediate both auto-regulatory and *trans*-regulatory translational control. In total, these findings reveal a repertoire of PABPC1 binding that is substantially broader than previously recognized with a corresponding potential to impact on and coordinate post-transcriptional controls critical to a broad array of cellular functions.

Contributions:

The contents of this section were generated by in collaboration with Hemant Kini. I performed all computational analyses on experimental data generated by Hemant Kini. We contributed equally to the drafting of the manuscript.

3.1 INTRODUCTION

The biogenesis of eukaryotic messenger RNAs (mRNAs) is tightly linked to the post-transcriptional addition of polyadenylate (poly(A)) tails to their 3' ends. These poly(A) tails contribute to regulation of mRNA transcription, transport, stability, and translation [42, 145]. PolyA tail-dependent functions are mediated in large part *via* the association of one or more poly(A) binding proteins (PABPs). In mammals, there are six defined PABP isoforms; a single nuclear isoform, PABPN1, that impacts on the addition of poly(A) tails in the nucleus and five cytoplasmic PABPs, ePAB, PABPC1, PABPC2, PABPC4, and PABPC5 that are thought to play roles in regulating mRNA stability and translation in the cytoplasm [42-44]. The overall structures and RNA binding specificities of the five cytoplasmic PABPs are highly conserved [45, 46]. They each contain four RNA Recognition Motifs (RRMs). RRMs 1 and 2 are primarily responsible for the high affinity binding to homopolymeric adenosines (K_d = 1.8 nM) [47], while RRMs 3 and 4 can bind to non-homopolymeric AU sequences (K_d= 2.9 nM) [47]. The levels of functional specificity and/or redundancy of the mammalian cytoplasmic PABPs remain unexplored.

PABPC1 is the major cytoplasmic PABP isoform in adult mouse somatic cells and is abundantly expressed in all tissues [48]. The interaction of PABPC1 with the poly(A) tails is well documented and defined in multiple contexts [42, 49]. The corresponding functions of the PABPC1/poly(A) tail complex are primarily mediated in pathways of mRNA stabilization and translation enhancement [50-52]. These functions may be linked to the interactions of PABPC1 with the 5' cap-binding complex *via* heterodimerization with eIF4G [53, 54]. Limited evidence points to additional binding sites and functions for PABPC1 within the eukaryotic mRNA transcriptome. For example, PABPC1 has been shown to bind to an A-rich element in the 5' untranslated region (UTR) of its own mRNA (mouse and human), establishing an auto-regulatory translational control circuit [41, 56, 57]. A recent study in *Saccharomyces cerevisiae* using a photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation approach (PAR-CLIP) demonstrated *in vivo* binding of yeast poly(A) binding protein Pab1 to AU-rich elements in

mRNAs [59], including binding to the efficiency element (UAUAUA) of the yeast polyadenylation signal [60, 61]. The impact of Pab1 binding to the polyadenylation efficiency element in yeast remains undefined, as does any generalization of these findings to higher eukaryotic organisms.

The extent to which PABPC1 binds to genomically encoded sequences in the mammalian transcriptome remains undetermined. The presence of such interactions could have broad implications to the understanding of post-transcriptional gene regulation. To address this gap, we comprehensively mapped PABPC1 binding to sites throughout the mouse transcriptome. This analysis revealed robust PABPC1 occupancy within the 3' untranslated region (3' UTR) that is predominantly localized to the canonical polyadenylation signal (PAS). A distinct set of PABPC1 interactions, lacking a defined binding site motif, were mapped to 5' and 3' boundaries of the open reading frame (ORF), exemplified most clearly in the replication-dependent histone mRNAs. A third, and more restricted subset of PABPC1 binding sites, was identified at AU-rich sites within the 5' UTRs of a small group of mRNAs and was demonstrated to impact on translation regulation. These studies substantially expand the known repertoire of PABPC1 interactions within the eukaryotic transcriptome and link a subset of these interactions to pathways of posttranscriptional control.

3.2 RESULTS

3.2.1 CLIP-seq identifies genomically encoded PABPC1 binding sites

We performed crosslinking immunoprecipitation coupled to high-throughput sequencing (CLIP-seq) to map PABPC1 binding sites within the transcriptome of mouse erythroleukemia (MEL) cells (see work-flow; Figure 3.1A). PABPC1 RNP complexes were captured by *in vivo* UV-crosslinking, followed by limited RNase I digestion, ^{32}P-labeling of RNA in the complexes, and immunoprecipitation with an isotype-specific anti-PABPC1 antibody. The immunoprecipitated PABPC1 RNP complexes were resolved on an SDS-PAGE gel and complexes migrating in close

proximity to the PABPC1 band (blue line, Figure 3.1B) were excised for analysis. The slower running complexes were excluded from library preparation as they were assumed to represent PABPC1 multimers bound to poly(A) (red line, Figure 3.1B). RNA fragments were isolated from the PABPC1 RNP complexes and used as templates for the construction of high-throughput sequencing libraries (see Methods).

Sequencing of the libraries generated from the PABPC1-bound RNA fragments yielded 14.8 million unique sequences across three biological replicates (Table 3.1). The mean size of the unique sequences protected from the RNase I treatment was 24 nucleotides (nts). These unique sequences were mapped to the mouse genome (mm10) with Novoalign, resulting in 7.5 million uniquely mapping CLIP tags that were used for downstream analysis (see Methods) (Table 3.1). We first examined the correlation of CLIP tags per gene between biological replicates and found a high level of reproducibly between experiments (Spearman correlation coefficient; R > 0.96 for all comparisons) (Figure 3.2A). This consistency between replicates allowed us to merge the biological samples for subsequent analyses. We also determined the correlation between PABPC1 CLIP-seq and mRNA-seq data that we generated from MEL cells (see Methods; Figure 3.2B). The relatively high correlation coefficient between this CLIP-seq and mRNA-seq comparison (Spearman correlation coefficient; $R^2 = 0.79$) suggested that the PABPC1 may recognize genomically encoded sequences shared by most mRNAs in the transcriptome. Overall, the high reproducibility of CLIP-seq replicates supported successful enrichment of PABPC1 bound fragments and warranted further investigation.

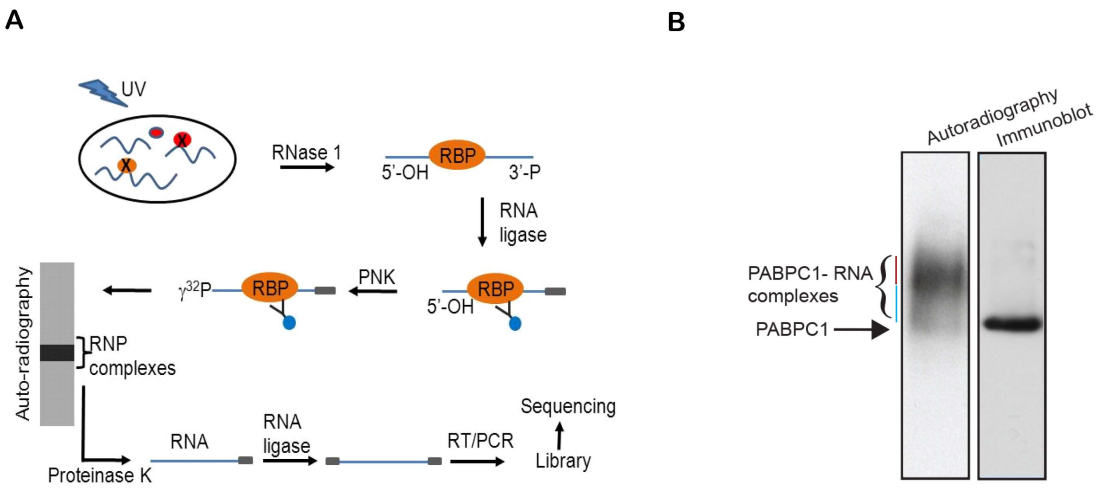


Figure 3.1 Overview of PABPC1 CLIP-seq. (A) Schematic of PABPC1 CLIP workflow showing immunoprecipitation and library preparation of PABPC1 bound RNAs for sequencing (see Methods). (B) Isolation and ^{32}P-labeled PABPC1-RNP complexes. Autoradiograph (left) and western blot (right) with antibody against PABPC1.

Table 3.1 Summary of PABPC1 CLIP-seq libraries.

PABPC CLIP-seq	Unique Tags	Tag Length	Uniquely Aligned Tags	Deletions	Deletions (%)	CIMS Sites	CIMS sites (P<.001)	5' UTR Clusters (mFDR <0.01)
rep1	3,715,048	22.93	1,817,709	135,014	7.43%	72,099	**11,907**	**2,824**
rep2	6,767,526	24.70	3,575,653	279,770	7.82%	127,748		
rep3	4,350,106	23.21	2,121,739	144,472	6.81%	74,005		
Merged	**14,832,680**	**23.82**	**7,515,101**	**559,256**	**7.44%**	**213,817**		

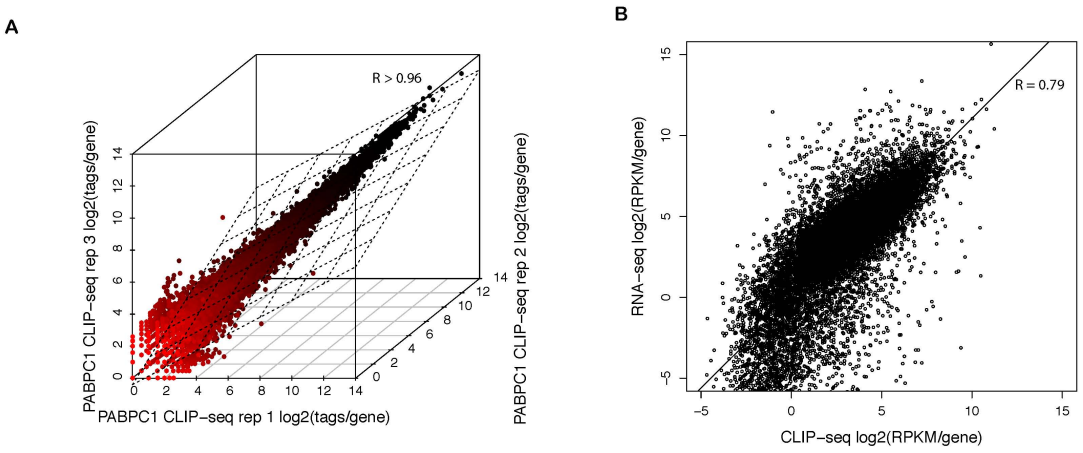


Figure 3.2 PABPC1 CLIP-seq libraries are reproducible. (A) Correlation of PABPC1 CLIP-seq replicates. PABPC1 CLIP-seq tags per gene are plotted for three independent biological replicates (Spearman correlation coefficient, R > 0.96 for all comparisons). (B) Correlation of PABPC1 CLIP-seq and RNA-seq from MEL cells. RPKM per gene is plotted for CLIP-seq and RNA-seq (Spearman correlation coefficient, R = 0.792).

3.2.2 PABPC1 binds predominantly to the 3' UTR of mRNAs

We next examined the distribution of PABPC1 CLIP tags across the mouse transcriptome. The great majority (73.89%) of CLIP tags mapped to the 3' UTR, with the next highest amount of tags (19.63%) mapping to the coding sequence (CDS) (Figure 3.3A). Since the average genomic length of 3' UTRs is shorter than that of the CDS, this distribution indicates a strong enrichment of PABPC1 binding in 3'UTRs. The remainder of the CLIP tags mapped to annotated 5' UTRs, introns, long intergenic non-coding RNAs (lincRNAs), and microRNAs (miRNAs). A meta-analysis of PABPC1 CLIP tags across mature mRNA transcripts, demonstrated a marked enrichment of CLIP tag density in proximity to annotated 3' termini of mRNAs (Figure 3.3B). This is in agreement with the majority of CLIP tags mapping to the 3' UTR. Visual inspection of specific mRNAs revealed numerous CLIP tags clustering along the 3' UTR with the most prominent peak occurring close to the 3' termini (examples in Figure 3.3C). Together, these data indicate that PABPC1 binds to genomically encoded sequences in numerous mRNAs, that most binding events occur in the 3'UTR, and that the preponderant localization of the PABPC1 binding occurs in close proximity to the 3' terminus of mRNAs.

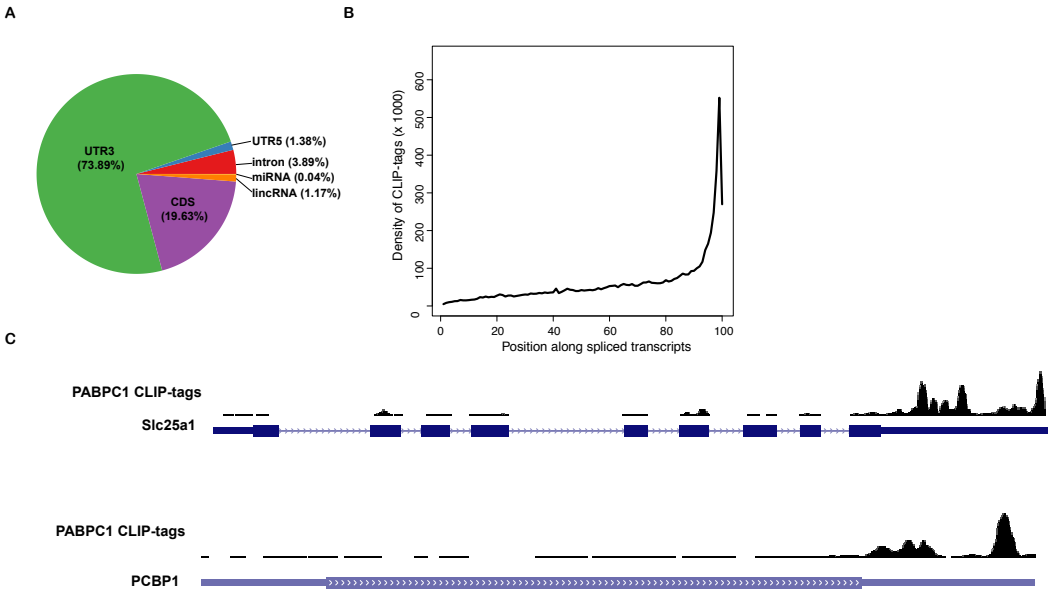


Figure 3.3 Distribution of PABPC1 CLIP tags. (A) Pie chart of the distribution of PABPC1 CLIP-tags within the transcriptome. (B) Relative distribution of PABPC1 CLIP tags along spliced mRNA transcripts. Gencode mRNAs were binned into 100 evenly sized regions and the coverage at each bin was used to create a composite profile. (C) Screenshots of the UCSC genome browser for two representative mRNAs (*Slc25a1* and *Pcbp1*), showing distribution of PABPC1 CLIP-tags along the length of the primary transcript. Note: *Pcbp1* is encoded by an intronless gene.

3.2.3 PABPC1 binding is enriched at the termini of 3' UTR

To enable closer inspection of PABPC1 binding sites, we mapped direct binding events at single-nucleotide resolution by crosslink induced mutation site (CIMS) analysis [146, 147]. This analysis takes advantage of the propensity for reverse transcriptase to skip nucleotides with protein adducts that remain after proteinase K treatment of RNP complexes. In agreement with previous studies, we found that deletions, but not insertions or substitutions were enriched within the body of CLIP tags (Figures 3.4A-C). CIMS analysis of PABPC1 CLIP tags identified 11,907 significant (False Discovery Rate (FDR) < 0.001) direct binding sites within the mouse genome (Table 3.1). Within the transcriptome, 86% of CIMS sites were located in 3' UTRs, 9% in the CDS, and the remaining 5% distributed across other regions (Figure 3.4D). To examine the distribution of CIMS sites in more detail, regions of mature mRNAs (5' UTR, CDS, and 3' UTR)

were binned into 100 discrete units, and CIMS coverage across each bin was calculated (Figure 3.5). We did not observe any positional enrichment within the 5' UTR, whereas an increase in binding events was observed towards the 3' end of the CDS leading into the beginning of the 3' UTR. While numerous binding sites were distributed throughout the 3'UTR, the most robust sites of enrichment for PABPC1 binding localized to the terminal segments of 3' UTRs. This distribution of CIMS sites is consistent with the distribution of CLIP tags toward the 3' terminus of mRNAs (Figure 3.3) and is indicative of direct PABPC1 binding to these regions.

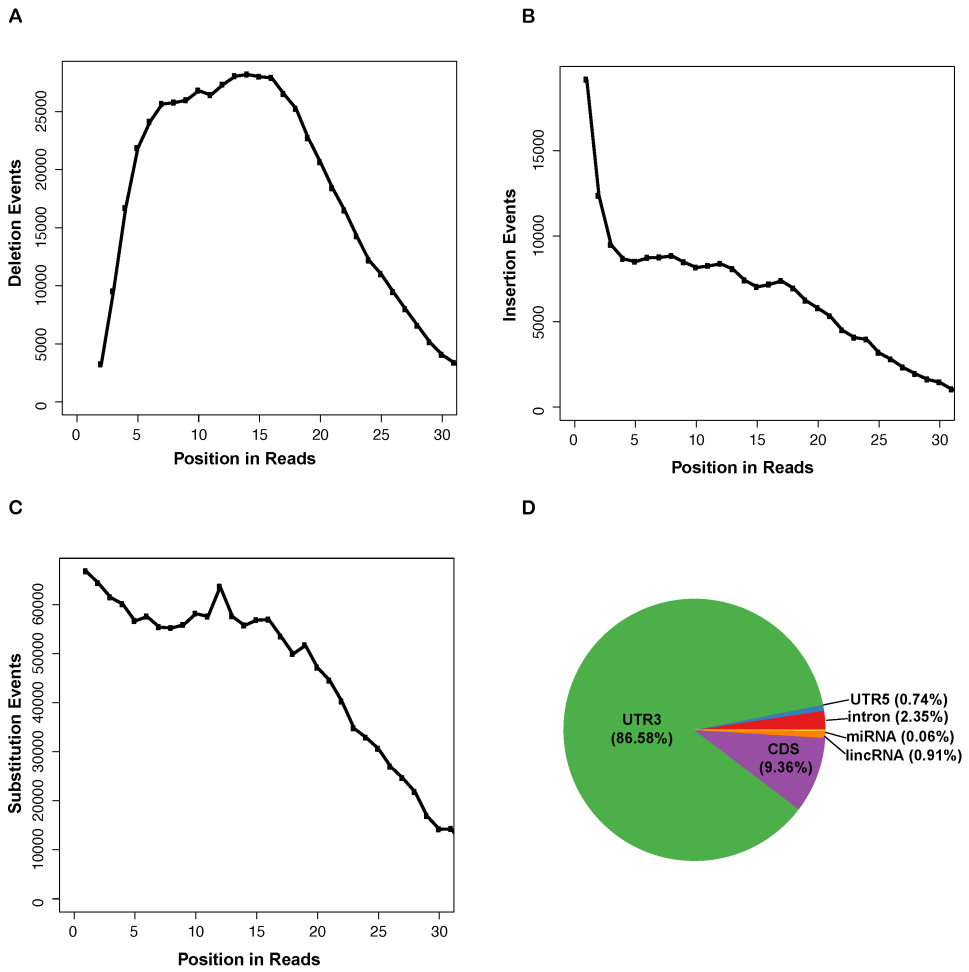


Figure 3.4 CIMS analysis of PABPC1 CLIP tags. (A) Absolute distribution of deletion events within the length of CLIP tags. (B) Absolute distribution of insertion events within the length of CLIP tags. (C) Absolute distribution of substitution events within the length of CLIP tags. (D) Distribution of CIMS sites in the mouse transcriptome.

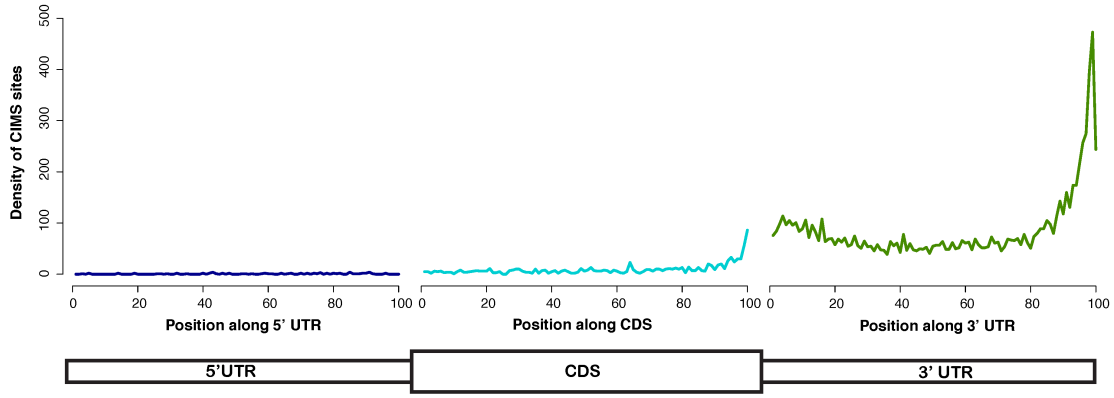


Figure 3.5 Relative distribution profile of CIMS sites along mRNAs. Position-specific coverage was calculated by parsing each region (UTRs and CDS) into 100 distinct bins and calculating CIMS coverage per bin.

3.2.4 PABPC1 binding sites are enriched for A/U-rich and A-rich motifs

We next determined the binding site sequence preference for PABPC1 within the mRNA population using multiple approaches. First, we analyzed the sequence content at each position surrounding CIMS sites. To do this, we anchored the analysis at CIMS sites and identified the base composition at each position +/- 10 nt from the CIMS site (Figure 3.6A). This approach revealed a strong preference for Adenosines interspersed with less frequent Uridines (denoted as T's on the logo). Uridine was the most commonly cross-linked base (position 11), consistent with analyses of UV-induced cross-links for other RBPs [148] and most likely reflecting preferential formation of UV-induced cross-linking of proteins with Uridine over other ribonucleosides [149]. We also calculated the enrichment of hexanucleotide sequences in the region +/- 15nt from each CIMS sites relative to all mRNA sequences. We selected the top 20 most enriched hexanucleotides and created a position weight matrix and motif logo to represent the sequence content (Figure 3.6B). A strong enrichment for Adenosine and Uridine was observed using this approach. Finally, using a *de novo* motif discovery algorithm (MEME) [135], we identified an A/U-rich sequence that had a striking resemblance to the canonical mammalian polyadenylation signal sequence (AAUAAA) (Figure 3.6C). The presence of this abundant and conserved sequence element may overshadow the identification of other true motifs. Therefore, to search for

a secondary motif, we eliminated all sequences that contained the top 10 mammalian PAS sequences (corresponding to 55% of all CIMS sequences) and re-ran the MEME analysis [150]. This secondary search revealed a purely A-rich sequence which was derived from ~170 CIMS sites (Figure 3.6D). Together, these orthogonal approaches led us to conclude that PABPC1 binds directly to both the PAS-like as well as to purely A-rich sequences within the mammalian transcriptome.

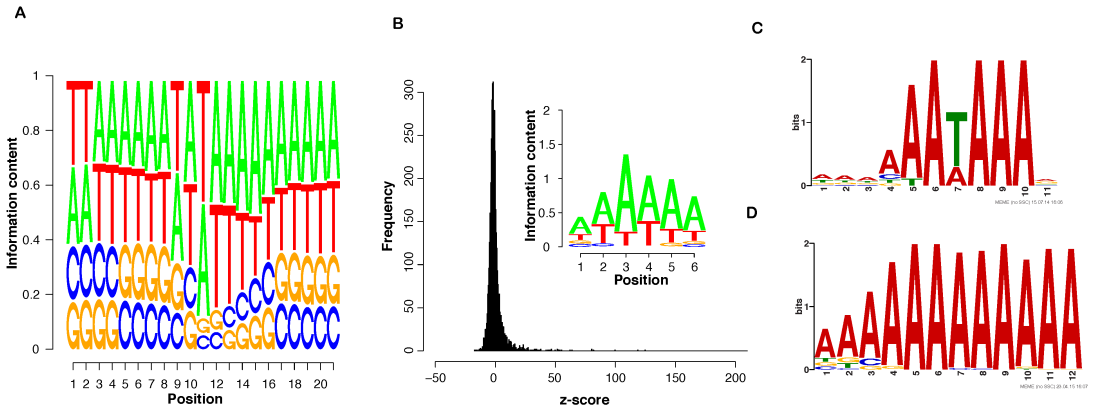


Figure 3.6 Motif analysis of PABPC1 CIMS sites. (A) Logo representing the average nucleotide sequence +/- 10 nt proximal to the CIMS sites (position 11 represents the CIMS site). (B) Z-score distribution of hexanucleotide analysis of CIMS sites (+/- 15 nt) and a logo representing the 20 most enriched hexanucleotides. (C) Motif logo uncovered by MEME analysis of CIMS sites +/- 15 nt flanking sequence. (D) Motif logo uncovered by MEME analysis on CIMS sites +/- 15 nt flanking sequence after removing all possible PAS signal sequences.

3.2.5 PABPC1 binds directly to the cleavage and polyadenylation signal

Due to the predominant binding of PABPC1 to 3' terminal poly(A) tails, we considered the possibility that the observation of enriched binding at the PAS might reflect 'bleed-over' from canonical poly(A) tail binding. Positional analysis of CIMS sites revealed that a preponderance of the PABPC1 CIMS sites mapped 20-25nt upstream of the annotated 3' terminus of mRNAs (Figure 3.7A). This location coincides precisely within the positioning of the PAS and argues against bleed-over from the poly(A) tail. Furthermore, alignment of the CIMS relative to the canonical PAS sequence (AAUAAA) revealed a sharp enrichment for CIMS sites specifically at this element (Figure 3.7B). To confirm that CIMS sites were localized to active PAS elements, we

used our mRNA-seq data generated from MEL cells to identify functional poly(A) addition sites in the MEL cell transcriptome (see Methods**).** We examined the distribution of active poly(A) addition sites relative to CIMS sites (Figure 3.7C). This analysis revealed that active poly(A) addition sites were located 20-25nt downstream of CIMS sites. These various approaches were internally consistent in demonstrating that PABPC1 binds directly to bona fide mRNA PAS elements. This binding to the PAS on mRNAs throughout the transcriptome is consistent with the high level of correlation between CLIP-seq and RNA-seq datasets and with the predominance of CLIP tags mapping to the 3' terminus of the 3'UTR (Figure 3.3). In summary, these data support the conclusion that PABPC1 binds directly to PAS elements and to genomically encoded A-rich mRNA sequences, in addition to its canonical role in binding to mRNA poly(A) tails.

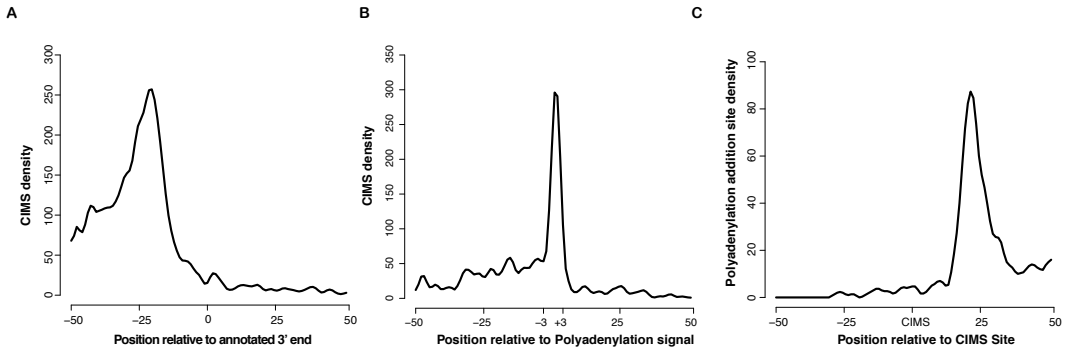


Figure 3.7 CIMS sites occur at the PAS. (A) Absolute distribution of CIMS relative to the end of annotated 3' UTRs. (B) Absolute distribution of CIMS sites relative to the PAS signal sequence (AAUAAA). (C) Absolute distribution of experimentally determined poly(A) addition sites relative to CIMS sites.

3.2.6 PABPC1 clusters are enriched in close proximity to the translation initiation and termination codons

As expected based on the correlation between mRNA abundance and PABPC1 CLIP tags, we found that the RNAs with the most CLIP tags were highly expressed mRNAs encoding the protein components of the ribosome and proteins that comprise the translation machinery. Surprisingly, we also observed that a number of replication-dependent histone mRNAs were represented among the top 1000 transcripts with highest CLIP tag density. Given that this class of

mRNAs is unique in lacking a PAS and poly(A) tail, we chose to examine the corresponding pattern of PABPC1 binding in detail. The distribution of CLIP tags across histone mRNAs (Figure 3.8A; as in Figure 3.3B) was prominently enriched at the 5' and 3' ends of transcripts. This distribution contrasts strongly with the 3' enrichment observed in the remainder of the mRNA transcriptome (compare Figs. 3.8A and 3.3B, examples in Figs. 3.8B and 3.3C). Detailed mapping of PABPC1 CLIP tags in replication-dependent histone mRNAs revealed that they were highly enriched over the translation initiation and termination sites (Figures 3.9A-B). Multiple analytic approaches failed to reveal any corresponding enriched primary sequence motif corresponding to these binding events. The pattern of binding within the replication-dependent histone mRNAs at start codons was similar to, although more sharply defined than that of the overall transcriptome (Figure 3.9C). PABPC1 binding in the vicinity of the stop codon of histone mRNAs similarly displayed a sharp peak (Figure 3.9B), whereas binding to the stop codon of all other detectable mRNAs in general peaked over the stop codon and remained high throughout the 3' UTR (Figure 3.9D). This difference in the contour of the CLIP tag mapping to the stop codon may reflect, at least in part, the high frequency of PABPC1 binding at the PAS in polyadenylated mRNAs. Together, these results suggest that PABPC1 interacts with the translation initiation and termination sites in a poly(A) and PAS-independent fashion.

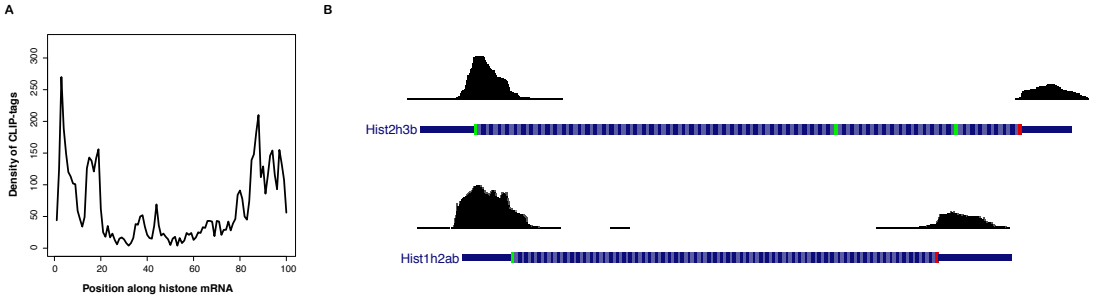


Figure 3.8 PABPC1 binds to histone mRNAs. (A) Distribution of PABPC1 CLIP tags along histone mRNA transcripts. Histone mRNAs were binned into 100 evenly sized regions and the CLIP-tag coverage in each bin was used to create a composite profile. (B) Screenshots from the UCSC genome browser for two representative histone genes showing CLIP tags proximal the 5' and 3' end of the CDS. Green and Red boxes indicate annotated start and stop codons, respectively.

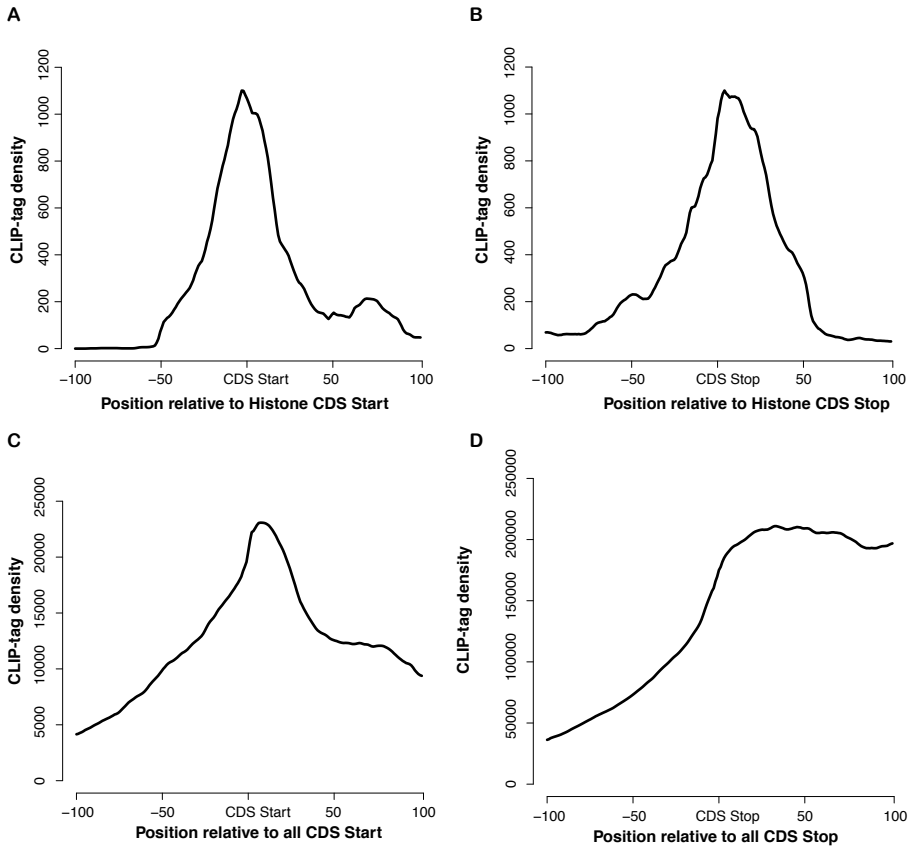


Figure 3.9 PABPC1 CLIP tags are enriched at start and stop codons. (A) Absolute distribution of CLIP tags proximal to the start codon of histone genes. (B) Absolute distribution of CLIP tags proximal to the stop codon of histone genes. (C) Absolute distribution of CLIP tags proximal to the start codon of all mRNAs. (D) Absolute distribution of CLIP tags proximal to the stop codon of all mRNAs.

3.2.7 PABPC1 binds to A-rich sequences within a subset of 5' UTRs.

Our initial analysis revealed a small subset of PABPC1 CLIP tags localized within 5' UTRs (Figure 3.3A). Interestingly, we found a relatively low correlation between number of CLIP-seq tags in 5' UTRs and mRNA abundance (mRNA-seq) (Figure 3.10, Spearman correlation coefficient; R = 0.356). This low correlation suggested that the binding of PABPC1 to the 5' UTR is heterogeneous across the population of mRNAs in MEL cells and occurs at determinants that are specific to subset(s) of transcripts. The number of CIMS sites within 5' UTRs (Figure 3.4D; 79 sites across 39 genes) was insufficient to identify a 5' UTR specific motif. However, analysis of

the 5'UTR CLIP tags by a low stringency approach (Pycioclip implementation of the modified false discover rate (mFDR) approach [151]), identified ~ 2,800 PABPC1 CLIP-tag clusters located in 5' UTRs (see Methods). Although 5' UTRs are generally G-C rich [152], MEME analysis of PABPC1 5' UTR cluster sites revealed an A-rich sequence motif that mapped to ~ 300 unique 5' UTR clusters (Figure 3.11A). We examined the correlation of CLIP tags from A-rich motif containing 5' UTRs with the mRNA-seq dataset and found an even weaker correlation (Figure 3.11B, Spearman correlation coefficient; R = 0.186) than what was observed for all 5' UTR CLIP tags (Fig. 4A). This lower correlation coefficient suggests that PABPC1 interactions with A-rich motifs in the 5' UTR are further uncoupled from mRNA steady state expression levels. A gene ontology analysis (DAVID; [137]) on this subset of mRNAs revealed enrichment for gene function terms involved in the regulation of transcription, DNA binding, nuclear processes, and cell cycle control (Figure 3.11C). Together, these results suggest that PABPC1 may coordinately regulate mRNAs involved in these basic cellular processes through binding to a shared 5' UTR motif.

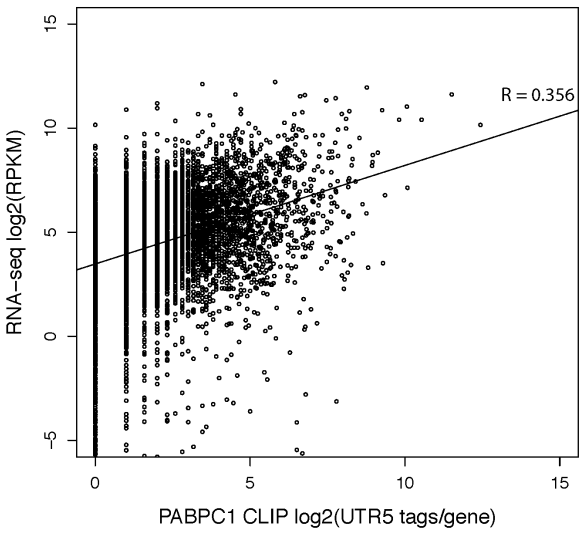
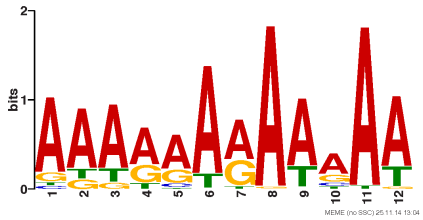
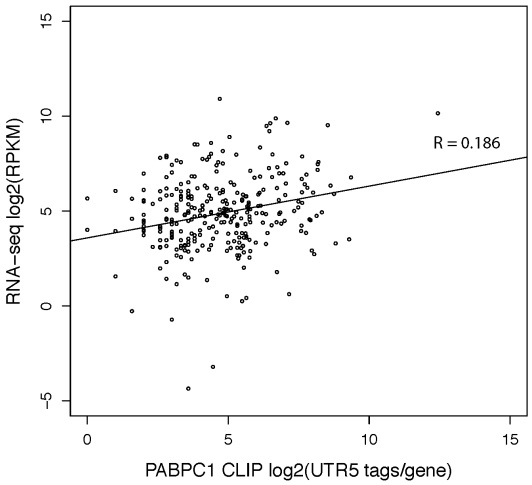


Figure 3.10 Correlation analysis of PABPC1 CLIP tags in 5' UTRs of protein-coding transcripts and mRNA-seq RPKM per gene values (Spearman correlation coefficient, R = 0.356).

A



B



C

Category	Term	PValue	Bonferroni	Benjamini	FDR
Biological Process	transcription	2.85E-07	3.88E-04	3.88E-04	4.68E-04
	regulation of transcription	2.24E-06	3.04E-03	1.52E-03	3.68E-03
	cell cycle	4.85E-05	6.39E-02	2.18E-02	7.96E-02
Molecular Function	DNA binding	2.96E-09	9.49E-07	9.49E-07	3.99E-06
	zinc ion binding	1.06E-06	3.41E-04	1.71E-04	1.43E-03
	transition metal ion binding	2.61E-06	8.37E-04	2.79E-04	3.52E-03
	RNA binding	4.15E-05	1.32E-02	3.32E-03	5.60E-02
Cellular Compartment	nuclear lumen	7.60E-07	1.72E-04	1.72E-04	9.72E-04
	intracellular organelle lumen	1.15E-05	2.60E-03	1.30E-03	1.47E-02
	organelle lumen	1.21E-05	2.74E-03	9.13E-04	1.55E-02
	membrane-enclosed lumen	2.29E-05	5.16E-03	1.29E-03	2.93E-02
	nucleolus	5.95E-05	1.34E-02	2.69E-03	7.61E-02

Figure 3.11 PABPC1 binds to A-rich motifs in the 5’ UTR of mRNAs. (A) Motif logo uncovered by MEME analysis of PABPC1 CLIP-tag clusters in the 5’ UTRs of mRNAs. (B) Correlation analysis of PABPC1 CLIP tags in 5’ UTRs of transcripts containing the A-rich motif (A) and RNA-seq RPKM per gene values (Spearman correlation coefficient, R = 0.186). (C) Gene ontology analysis of genes with A-rich motifs within CLIP-tag clusters in the 5’ UTR.

Interestingly, we found that the top A-rich PABPC1 binding site in the 5’ UTR corresponds to *Pabpc1* mRNA (Figure 3.12A). This is of note, because it has been previously reported that PABPC1 represses the translation of its own mRNA *via* binding to a 5’UTR A-rich determinant [153, 154]. Among other transcripts with significant PABPC1 binding to the A-rich sequence in the 5’ UTR were, cell cycle control protein Cyclin D2 (*Ccnd2*) (Figure 3.12B), Scaffold Attachment Factor B (*Safb*) an RNA binding protein that impacts on both transcription and splicing [155], and Adenosylmethionine Decarboxylase 1 (*Amd1*), a protein associated with cell and tumor growth [156], metabolism and obesity [157]. These observations led us to

hypothesize that PABPC1 binds to an A-rich determinant within the 5' UTR of a subset of mRNAs and this binding may be of particular importance to the regulation of their translation.

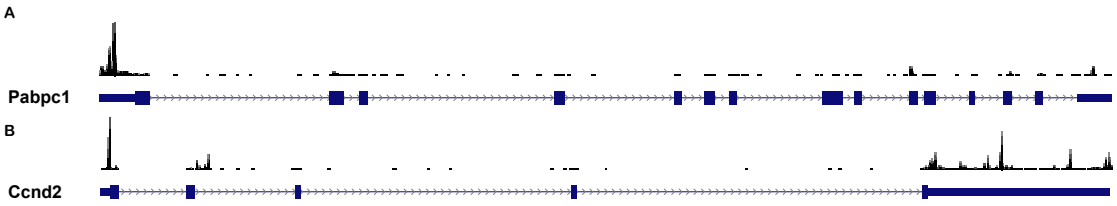


Figure 3.12 Screenshot of PABPC1 CLIP tags from the UCSC genome browser for two transcripts with PABPC1 binding within their 5' UTRs (*Pabpc1* (top) and *Ccnd2* (bottom)).

3.2.8 PABPC1 auto-regulates its expression by binding to an A-rich element

PABPC1 was previously found to repress its own translation by binding to a 5' UTR A-rich determinant [56, 158]. Interestingly, we identified two distinct clusters of PABPC1 binding sites in the *Pabpc1* mRNA 5' UTR (Figure 3.13A, red and green bars, respectively), a more 5' cluster, overlapping with the previously identified A-rich element ('5' cluster', red bar), and a second, larger cluster of unknown function ('3' cluster', green bar). To determine if these clusters have overlapping or unique roles in regulation of PABPC1 expression we cloned the intact PABPC1 5' UTR into a Firefly Luciferase reporter plasmid and separately inserted derivative 5'UTRs specifically lacking each of the two individual PABPC1 CLIP-tag clusters (Figure 3.13B). Each of these plasmids were transfected into NIH-3T3 cells. NIH-3T3 cells were chosen because they are more effectively transfected than MEL cells. Firefly Luciferase protein and RNA expression were quantified 48h post transfection (Figure 3.13C). Luciferase protein was significantly (C1-5' UTR vs Luc; $P < 0.01$; two-tailed T-test) repressed in the presence of native *Pabpc1* 5' UTR in the absence of an appreciable impact on mRNA accumulation (Figure 3.13C). This impact on expression was fully consistent with a mechanism of translation inhibition by PABPC1. Deletion of the 5' interaction site resulted in a significant increase in protein expression as compared to the intact *Pabpc1* 5' UTR (red cluster in Figure 3.13C, Mut1 vs. C1-5'UTR; $P < 0.001$; two-tailed T-test) while deletion of the 3' cluster had a statistically significant but marginal

impact on protein output (Mut 2 vs. C1-5'UTR; P < 0.001; two-tailed T-test). These data, based on luciferase reporter assay, suggest that the more 5' cluster mediates the translation repression activity of the intact 5' UTR, while the function of the more 3' cluster, if any, remains to be defined (see below).

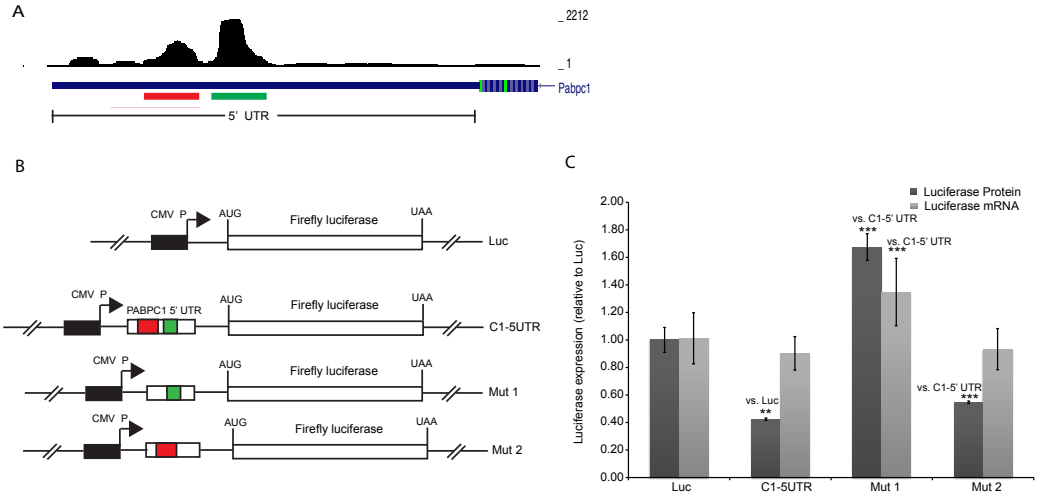


Figure 3.13 PABPC1 regulates the expression of its own mRNA. (A) Screenshot from UCSC genome browser of PABPC1 CLIP tags in the 5' UTR for *Pabpc1* mRNA. Red and Green bars highlight two major PABPC1 CLIP-tag clusters (5' and 3' clusters, respectively). (B) Insertion of the *Pabpc1* 5' UTR and three derivatives in an expression vector in frame with the Firefly luciferase ORF. Mutants represent deletion of either one or both of the CLIP-tag clusters denoted in (A). (C) Quantification of Firefly luciferase mRNA levels (qRT-PCR; light grey bars) and luciferase enzymatic activity levels as a proxy for protein abundance (luciferase assay; dark grey bars).

While PABPC1 has been reported to auto-regulate its protein expression by binding to the 5' UTR A-rich sequence, this was not clearly delineated from the luciferase assays as deletion of the 5' cluster also resulted in a significant (Mut 1 vs. C1-5' UTR; P < 0.05; two-tailed T-test) increase in luciferase reporter mRNA expression (Figure 3.13C). Also of interest, we found that the 3' cluster, whose deletion did not impact upon the reporter expression (Figure 3.13C), overlapped the start of a predicted upstream open reading frame (uORF) and co-localized with initiating ribosomes as determined by ribosome profiling with Harringtonin-treated mouse ES cells

[159] (Figure 3.14). Therefore, this region may regulate more complex translational control mechanisms not evident from these reporter assays.

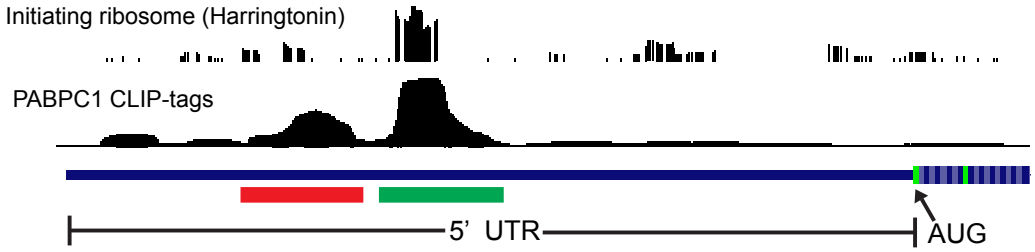


Figure 3.14 PABPC1 and initiation ribosomes bind to the same region of the *Pabpc1* 5' UTR. UCSC genome browser screenshot showing the region of overlapping PABPC1 CLIP tags and initiating ribosome tags within the *Pabpc1* 5' UTR. Arrow denotes the canonical start codon (AUG).

To address these discrepancies, we chose to examine the *in vivo* function of *Pabpc1* 5' UTR binding clusters by ablating them in cell lines with CRISPR/Cas9 mediated site-directed deletion. Two guide RNAs (gRNAs) corresponding to sites flanking the *Pabpc1* 5' UTR binding clusters were cloned into separate vectors expressing the Cas9 nuclease. These two vectors were co-transfected into the mouse myoblast cell line C2C12 and Puromycin-resistant clones containing a heterozygous deletion of this region were identified (Figure 3.15A). Analysis of *Pabpc1* 5' UTR Mut +/- cells revealed that deletion of the *Pabpc1* 5' UTR clusters resulted in a ~ 2 fold ($P < 0.01$; two-tailed T-test) increase in PABPC1 protein levels in the absence of an alteration in steady state mRNA levels (Figures 3.15B-C). These *in vivo* results strongly support the model in which PABPC1 binding within the 5' UTR of its encoding mRNA is critical for the homeostasis of PABPC1 protein expression. Deletion of the PABPC1 binding site in the 5' UTR results in increased levels of PABPC1 protein expression *in vivo* in the absence of an alteration in mRNA levels.

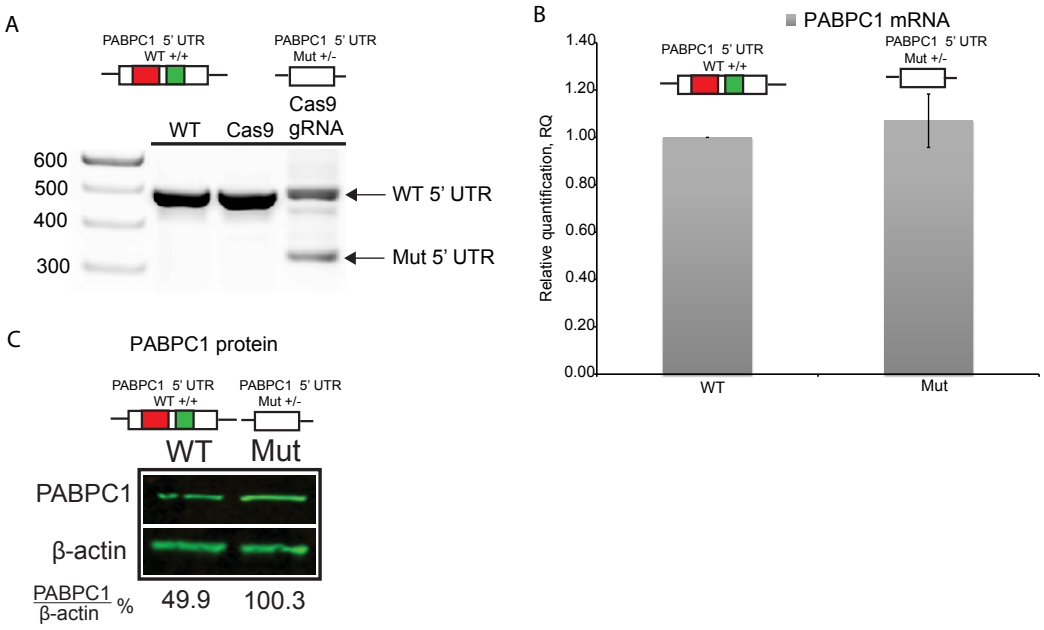


Figure 3.15 CRISPR analysis of *Pabpc1* 5' UTR. (A) Agarose gel of genomic DNA PCR showing *Pabpc1* 5' UTR region for WT cells (untransfected C2C12 cells), cells transfected with a vector expressing Cas9 without guide RNAs (Cas9), and cells transfected with vectors expressing Cas9 and gRNAs targeting sites flanking both of the PABPC1 CLIP-tag clusters (Cas9/gRNA). Upper and lower arrows denote WT and mutant loci respectively. (B) Quantification of *Pabpc1* mRNA levels (qPCR) for mutant clones relative to WT cells. (C) PABPC1 immunoblot of WT C2C12 cells and mutant clones. PABPC1 levels were quantified by densitometry and normalized to b-actin. ** = $P < 0.01$, *** = $P < 0.001$; Two-tailed T-test.

3.2.9 PABPC1 inhibits synthesis by binding to 5' UTR A-rich elements

We next sought to determine if PABPC1 binding to 5' UTR A-rich elements mediated regulatory control over additional mRNAs. Three mRNAs with prominent PABPC1 binding clusters at A-rich sites within their 5' UTR were chosen for study; *Safb*, *Amd1*, and *Ccnd2* (Figures 3.16A, C, and E). The full 5' UTR of each of these mRNAs was cloned into the Firefly Luciferase reporter plasmid and its impact was compared with the corresponding 5' UTRs lacking the PABPC1 binding site Figures 3.16B, D, and F). Deletion of the PABPC1 binding region from both the *Safb* and *Amd1* 5' UTRs significantly enhanced luciferase expression levels (*Amd1*: $P < 0.05$ and *Safb*: $P < 0.01$; two-tailed T-tests) without altering corresponding mRNA levels (Figure 3.17). A significant ($P < 0.01$; two-tailed T-test) increase in luciferase activity was also observed upon deletion of the PABPC1 binding site from the *Ccnd2* 5' UTR (Figure 3.17) although in this

case there was a corresponding increase in mRNA levels ($P < 0.05$; two-tailed T-test). This increase in mRNA levels was similar to what was observed for deletion of the *Pabpc1* 5' cluster region (Figure 3.13C). These results demonstrate that PABPC1 binding to A-rich sites within the 5' UTR of specific mRNAs can repress protein expression by repressing mRNA levels and/or by impeding effective translation. Thus, PABPC1 is involved in post-transcriptional control of gene expression in mammalian cells *via* an array of mechanistic pathways.

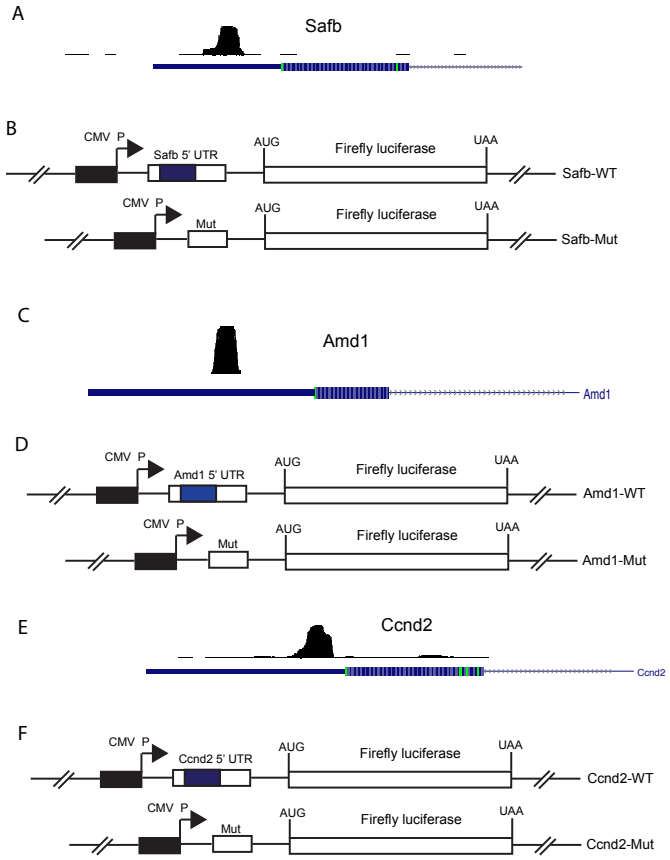


Figure 3.16 Screenshots and schematic of PABPC1 5' UTR targets. (A, C, and E) Screenshots from the UCSC genome browser of PABPC1 CLIP tags in the 5' UTR for *Safb* (A), *Amd1* (D), and *Ccnd2* (G) mRNA. (B, D, and F) The native *Safb* (B), *Amd1* (D), and *Ccnd2* (F) 5' UTRs (top) or mutant derivatives lacking the A-rich PABPC1 binding site cluster (bottom) were separately inserted in-frame with the Firefly luciferase ORF in a standard expression vector. The PABPC1 binding site, corresponding to the CLIP-tag cluster in (A), (C), and (E), is represented by the black rectangle within the 5' UTR.

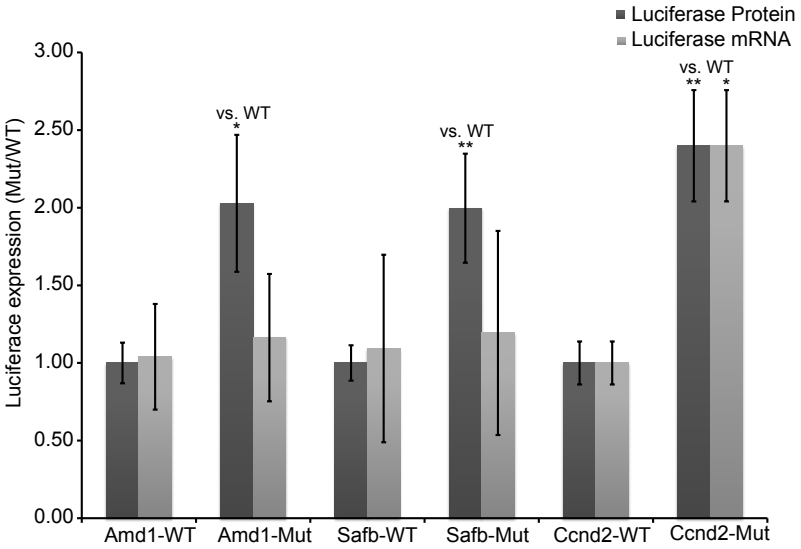


Figure 3.17 PABPC1 5' UTR binding sites regulate translation. Quantification of Firefly luciferase mRNA levels (qRT-PCR; light grey) and luciferase enzymatic activity levels (luciferase assay; dark grey) for the *Safb*, *Amd1*, and *Ccnd2* 5' UTR constructs. * = P < 0.05, ** = P < 0.01; Two-tailed T-test.

3.3 DISCUSSION

PABPC1 is an abundant cytoplasmic RNA-binding protein that is expressed in all somatic cells. The functions of PABPC1 are best understood in the context of its binding to the homopolymeric poly(A) tails of mRNAs. This PABPC1/poly(A) tail complex has been linked to pathways that control mRNA stability and translation activity and exerts significant impact on multiple cell functions [42, 145, 160]. In *Saccharomyces cerevisiae* deletion of its single poly(A) binding protein, Pab1p, is incompatible with cell viability [161] and in *Drosophila melanogaster*, homozygosity for P-element disruption of the cytoplasmic PABP gene results in embryonic lethality [162]. The impact of PABPC1 depletion or ablation in mammalian cells remains undefined.

Limited *in vitro* studies suggest that PABPC1 can bind to mRNAs at sites other than the poly(A) tail [47]. *In vivo* analysis in yeast provided further evidence that Pab1p binds to genomically encoded A and A/U rich sequences in mRNAs [59, 61]. In the present report, we performed CLIP-seq on PABPC1 in MEL cells with the goal of revealing the extent and role of PABPC1 binding to genomically encoded sequences in mammalian. These studies reveal that PABPC1 binds to complex sequences within different regions of annotated mRNAs and that subsets of these interactions have a defined impact on gene expression.

3.3.1 PABPC1 binds to the PAS of mRNAs throughout the mammalian transcriptome

A key observation from our study is that the majority of the PABPC1 CLIP tags cluster within mRNA 3' UTRs (73.89%, Figure 3.3A). This is not surprising as proteins that bind to the CDS are generally susceptible to displacement by the elongating ribosome and stably assembled RNP complexes are preferentially localized to the 3' UTR 'sanctuary' [163, 164]. As PABP's are well characterized for their strong association to mRNA poly(A) tails it was necessary to rigorously demonstrate that the enrichment within 3' UTRs reflected direct binding rather than 'bleed over' from binding to the adjoining poly(A) tails. Mapping of PABPC1 binding at single nucleotide resolution by a CIMS analysis (Figures 3.4 – 3.7) unambiguously identified that PABPC1 binds directly within the mRNA 3' UTR's (Figure 3.5). Coupled with two orthogonal approaches, we were able to further determine that the majority of these binding interactions are localized to the canonical cleavage and polyadenylation signal sequence (AAUAAA) (Figures 3.6-3.7). These observations are consistent with the prior analysis of Pab1p PAR-CLIP studies in yeast in which binding was mapped to the AU-rich efficiency element within the 3' UTR region [59]. These results lead us to conclude that binding of cytoplasmic PABPs to polyadenylation elements has been conserved from yeast to mammalian cells.

While the function(s) of non-poly(A) tail PABP RNP complexes in the cytoplasmic compartment remains unclear, related binding activities have been functionally linked to post-

transcriptional pathways of gene regulation. For example, the cytoplasmic polyadenylation element binding (CPEB) protein binds to an A-rich element (CPE) within the 3' UTR where it recruits the cleavage and specificity factor (CPSF) to the PAS with consequent cytoplasmic extension of the poly(A) tail and translational enhancement [165]. CPEB can also recruit proteins such as Maskin to regulate mRNA translation in Xenopus oocytes [165] and mouse hippocampus [166]. It is plausible that PABPC1, once bound to the PAS, can recruit other trans-acting factors that modulate translation. Furthermore, these mechanisms might reflect direct actions on translation or mRNA stability, or alternatively, the impact may be indirect, reflecting an impact on the length and/or function of the poly(A) tail. Based on our mapping data, these and related models can now be fully explored.

3.3.2 PABPC1 binds in close proximity to the translation initiation and termination codons

The mapping of CLIP tags within the MEL cell transcriptome revealed robust binding to replication-dependent histone mRNAs. These mRNAs are unique amongst polymerase II transcribed mRNAs in that they lack PAS elements and poly(A) tails. Analysis of the histone mRNAs thus allowed us to focus on PABPC1 interactions in the absence of the predominant poly(A) tail and PAS binding activities. Intriguingly, the CLIP tags within histone mRNAs localized to the sites of translation initiation and termination (Figures 3.8 – 3.9A-B). We observed similar enrichment for CLIP tags at the start codons throughout the transcriptome while the signal at stop codons was somewhat overshadowed in the bulk of mRNAs by PAS binding (Figures 3.9C-D). Importantly, the small number of CIMS sites and lack of any enriched sequence motif for PABPC1 binding at sites flanking ORFs, suggests that enrichment of CLIP tags bracketing the open reading frame may reflect indirect association of PABPC1. This indirect positioning of PABPC1 is consistent with the model proposed by others that PABPC1 remains associated with

the elongating ribosome during translation [167, 168]. Further study will be necessary to understand the role and functional consequences of PABPC1 binding to these regions.

3.3.3 PABPC1 binds to A-rich sites within the 5'UTR of a restricted subset of mRNAs with resultant post-transcriptional repression of gene expression

The binding of PABPC1 within 5' UTRs appears to be limited to a highly restricted subset of mRNAs (Figures 3.10-3.12). This specificity is indicated by the lack of correspondence between mRNAs bound in this region by PABPC1 and overall mRNA representation in the transcriptome (Figures 3.10 and 3.11B). MEME analysis of 5'UTR clusters revealed enrichment for a predominantly A-rich motif, consistent with the binding site preference of the PABPs (Figure 3.11A). Interestingly, the highest ranked PABPC1 binding target within this mRNA subset was Pabpc1 mRNA. PABPC1 has been previously reported to auto-regulate its own translation by binding to an A-rich domain within the 5' UTR [153]. This translational control domain is coincident with a prominent PABPC1 CLIP-tag cluster identified in the current study (5' cluster, highlighted in red, Figure 3.13A). Remarkably, this analysis also revealed an adjacent and even more prominent cluster of CLIP tags (3' cluster, highlighted in green, Fig. 5A) that did not impact on translation (Figure 3.13C, C1-5' UTR vs. Mut 2). Interestingly, this second PABPC1 binding region tracks with the positioning of the initiating ribosome in mouse ES cells as mapped by ribosomal profiling with Harringtonin treatment and was predicted to encode the start of a uORF [159] (Figure 3.14). Thus, this PABPC1 binding element within the 5' UTR may yet play a role in translational control not captured by the luciferase assay (Figure 3.13C).

To validate the *in vivo* function of *Pabpc1* 5' UTR clusters in translational control, we deleted the region of the *Pabpc1* 5' UTR spanning both of the PABPC1 CLIP-tag clusters *via* Crispr/Cas9 endonuclease targeting. The 2-fold increase in PABPC1 protein expression in cells heterozygous for the 5'UTR deletion in the absence of any alteration in mRNA stead state levels, confirmed that this region acts to auto-regulate *Pabpc1* translation (Figure 3.14C). Importantly,

PABPC1 overexpression has been associated with defective spermiogenesis in mice [169], deadenylation and translation inactivation in Xenopus oocytes, and with variations in cell cycle and apoptosis in certain leukemias [170]. Thus, this auto-regulatory feature of the *Pabpc1* 5' UTR mediates a regulatory pathway relevant to critical aspects of cell differentiation and proliferation.

The potential for PABPC1 to control gene expression was further extended by the analysis of additional mRNAs identified with 5'UTR PABPC1 CLIP-tag clusters. Deletion of these binding site regions enhanced the translation of reporter expressing *Safb* and *Amd1* mRNA 5' UTR's (Figure 3.17). Protein expression was also enhanced by similar deletion within the 5'UTR of the *Ccnd2* mRNA, although in this case there was a concomitant increase in the steady state mRNA levels (Figure 3.17). We note that deletion of the 5' cluster (Mut 1) in the PABPC1-5' UTR (Figure 3.13C) also enhanced mRNA levels, although to a lesser extent than the corresponding protein expression. These data underline the potential for the 5' UTR binding of PABPC1 to impact on a variety of mechanisms that repress gene expression, including both mRNA stability as well as translational control. The relative importance of each of these pathways may reflect specifics of the binding site, including interaction with other trans-factors and/or adoption of specific RNA secondary structures. Overall, we reveal that PABPC1 regulates a repertoire of gene regulatory pathways and establish a foundation for the exploration of additional targets and cellular functions mediated by PABPC1 binding to genomic mRNA sequences.

3.5 MATERIALS AND METHODS

Cell culture and CLIP-seq analysis

MEL and NIH-3T3 cells were grown under standard conditions in minimal essential medium (MEM) and Dulbecco's modified Eagle medium (DMEM), respectively, supplemented with 10% (vol/vol) fetal bovine serum (FBS) and 1× antibiotic-antimycotic (Invitrogen, Carlsbad, CA). MEL cells were washed with Hanks' balanced salt solution (HBSS) and cross-linked with UV (400 mJ/cm^2) three times on ice. CLIP was performed according to previously published protocol [98, 171]. Briefly UV cross-linked MEL cells were lysed with 1x PMPG in the presence of RNase 1

(2.5 U, Promega, Madison, WI), DNAse I (Promega, Madison, WI) treated for 15 min. The lysates were ultra-centrifuged at 90,000g for 20 mins. Immunoprecipitation was performed with protein A Dynabeads coated with PABPC1 antibody (Abcam, Cambridge, MA). Following the wash steps radiolabeled 3' adaptor was ligated to the complexes on the beads using T4 RNA ligase (Thermo Scientific) for 16 h at 16^{o}C. The beads were then washed, treated with T4 polynucleotide kinase (NEB, Ipswich, MA), and the RNP complexes were eluted off of the beads. 90 % of the eluate were used for autoradiography and the remainder was used for immunoblotting. The RNP complexes were resolved on 4- 12 % NuPage gels (Invitrogen, Carlsbad, CA), transferred to nitrocellulose membrane, and then exposed to X-ray film. Using the X-ray film as a guide, the portion of the nitrocellulose membrane corresponding to PABPC1-RNA complexes was excised, Proteinase K (Roche, Basel, SUI) treated, and the RNA was Phenol extracted. The purified RNA was ligated to a 5' adaptor, amplified, and sequencing libraries were constructed. Libraries generated from biological triplicates were individually bar coded, pooled, and sequenced on an Illumina HiSeq 2000 platform at the University of Pennsylvania Next Generation Sequencing Core (NGSC).

CLIP-seq read processing and alignment

Adapter sequences (*GTGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCTGCTTG*) were removed from raw reads and only trimmed reads were used for downstream analysis. Trimmed reads from each individual replicate CLIP-seq experiment were collapsed and mapped to the mouse genome (mm10) with Novoalign (Novocraft, Selagnor, MYS) with the parameters –t 85 -l 15 –s 1 –o Native –r None. Replicate experiments were merged and only uniquely mapped reads were used for subsequent analysis.

CIMS and Cluster analysis

CIMS analysis was applied to identify single-nucleotide RBP-RNA interaction sites (as described; Moore et al. 2014). Briefly, deletion sites were extracted for each CLIP tag from

novoalign output and a negative binomial test was used to assess significance. Sites with FDR < 0.001 were used for downstream analysis. To identify significant CLIP-seq clusters we used Pyicoclip [151] with an mFDR < 0.01. Gencode annotation vM2 was used for all analyses.

mRNA-seq

mRNA-seq was performed as previously described [172] . Briefly, total RNA was purified from the MEL cell cultures (miRNeasy; Qiagen, Valencia, CA). Poly(A)+ RNA was isolated using oligo dT beads (Life Technologies, Frederick, MD). RNA was fragmented for 7 minutes using Fragmentation Reagent (Life Technologies, Waltham, MA). mRNA-seq libraries were then generated using the Illumina smRNA-seq kit (illumina, San Diego, CA). Reads were trimmed with Cutadapt, mapped with Tophat2, and gene expression was quantified using HTseq [173-175]. Custom python scripts were used to calculate RPKM.

Motif analysis

Motif analysis was carried out by aligning CIMS sites and extracting sequences +/- 10nt from each site. A custom script was used to create a position-weight matrix and used the R package SeqLogo to generate motif logos [176]. For hexanucleotide enrichment analysis, the equally sized regions in the exonic portion of the mRNA transcriptome was shuffled10 times and the prevalence of each hexanucleotide was calculated and compared to the abundance in +/- 15nt CIMS regions. A position weight matrix was created from the top 20 hexanucleotides. For *de novo* motif discovery, MEME was used with a maximum width of 12nt [135].

Active polyadenlyation addition site identification

To identify high confidence polyadenylation additions sites, we used a custom python script to filter raw mRNA-seq reads with at least 20 Adenines at the 3' end. We then removed these poly(A) stretches, mapped the remaining sequence to the mouse genome with Tophat2,

and calculated the density of 3' ends using bedtools genomecov [177]. Only sites with greater than 10 reads per million were considered bona-fide poly(A) sites.

Luciferase assays

5' UTR or defined variants were cloned into a Firefly luciferase vector. These constructs were transfected into NIH-3T3 cells in 12-Well plate using Turbofect transfection reagent (Thermo Scientific). After 48h Luciferase activity was measured using Dual Luciferase assay kit (Promega, Madison, WI) and the corresponding mRNA levels were quantified by qPCR.

CRISPR targeted deletion of PABPC1 5' UTR region

gRNA oligos (ATAAATGTGTGTTCCGAGCCCGG) and (TCGGTCTCGGCTGCTTCACCGGG) were designed using the Broad Institute CRISPR design tool (www.crispr.mit.edu). After restrictions digest with BbsI they were cloned into px330 vector and then transfected into C2C12 cells using Lipofectamine 3000 (Life Technologies, CA). After 72 h, purmomycin was added at 1 mg/ml to and colonies with targeted 5' UTR deletions were selected for gDNA PCR.

Quantitative Western blotting

Cells were lysed in radioimmunoprecipitation assay (RIPA) buffer, and the following primary and secondary antibodies were used: rabbit anti-PABPC1 (Abcam), rabbit anti-actin (Bethyl), and goat-anti-rabbit IgG (Licor). Blots were visualized and scanned with Odyssey scanner and software (Li-Cor Bioscience).

Re-analysis of Ribosome Profiling data

Ribosome profiling data from harringtonin-treated mouse embryonic stem cells (mESCs) were obtained from GSE30839. We processed the ribosome profiling data as previously

described [178]. Briefly, reads were trimmed for adapter sequence (*CTGTAGGCACCATCAATTCGTATGCCGTCTTCTGCTTGAA*), filtered by mapping to mouse ribosomal RNA sequences. Filtered reads were mapped to the mouse transcriptome and genome using TopHat2 [174, 179]. Only mapped reads with no mismatches were used for further analysis. Aminoacyl-tRNA sites were identified as previously described [178].

Availability of supporting data:

The data sets supporting the results of this article are available in the GEO repository, under accession number GSE69755.

Chapter 4: ISOLATION AND SEQUENCING OF AGO-BOUND RNAS

This section refers to work from:

- Silverman IM*, Gosai SJ*, Vrettos N, Foley SW, Berkowitz ND, Mourelatos Z, Gregory BD. Isolation and sequencing of AGO-bound RNAs reveals characteristics of mammalian stem-loop processing *in vivo*. In Prep

Abstract:

MicroRNA precursors (pre-miRNAs) are short hairpin RNAs that are rapidly processed into mature microRNAs (miRNAs) in the cytoplasm. Due to their low abundance in cells, sequencing-based studies of pre-miRNAs have been limited. We successfully enriched for and deep sequenced pre-miRNAs in human cells by capturing these RNAs during their interaction with Argonaute (Ago) proteins. Using this approach, we detected > 350 pre-miRNAs in human cells and > 250 pre-miRNAs in a reanalysis of a similar study in mouse cells. We uncovered widespread trimming and non-templated additions to 3' ends of pre-miRNAs and mature miRNAs. Additionally, we identified novel Ago2-cleaved pre-miRNAs and created an index for microRNA precursor processing efficiency. This analysis revealed a subset of pre-miRNAs that produce low levels of mature miRNAs despite abundant precursors, including an annotated miRNA in the 5' UTR of the DiGeorge syndrome critical region 8 (Dgcr8) mRNA transcript. This led us to search for other Ago-associated stem-loops originating from mRNA species, which identified hundreds of putative pre-miRNAs embedded within mRNA sequences in both the mouse and human transcriptomes. Intriguingly, we found that iron responsive elements in ferritin heavy and light chain mRNAs are processed into Ago-associated stem-loops in both mouse and humans but do not produce functional small RNAs. In summary, we provide a wealth of information on pre-miRNAs, and identified microRNA and microRNA-like elements in mRNAs.

Contributions:

The contents of this section were generated by in collaboration with Sager Gosai, Nicholas Vrettos, and Shawn Foley. I performed the initial experimental analysis with assistance from Nicholas Vrettos. Sager Gosai developed the computational pipelines and together we performed all bioinformatic analyses. Shawn Foley performed qPCR validation experiments. I drafted the manuscript with assistance from other authors.

4.1 INTRODUCTION

MicroRNAs (miRNAs) are ~22 nucleotide (nt) small RNAs (smRNAs) that function in post-transcriptional gene regulation to repress translation or promote degradation of target messenger RNAs (mRNAs) [17, 180]. Animal miRNAs are generated in a two-step process, whereby miRNA precursors (pre-miRNAs) are first cleaved from their primary transcripts by the action of the DGCR8/DROSHA microprocessor complex [67, 68]. Alternative pre-miRNA biogenesis pathways have been described that bypass the microprocessor, for example mirtron loci generate pre-miRNAs in a splicing-dependent, DROSHA-independent fashion [70, 71]. Pre-miRNAs are then transported into the cytoplasm by Exportin-5 for further processing [72, 73].

Cytoplasmic pre-miRNAs are then matured by the miRNA loading complex (miRLC) which is composed the type III endonuclease DICER, the double-stranded RNA-binding protein TRBP, and the miRNA effector protein Argonaute (AGO) [74-76]. DICER cleaves the pre-miRNA to reveal a ~22 nt miRNA duplex, consisting of the upstream miRNA (denoted as the 5p miRNA) and downstream miRNA (denoted as the 3p miRNA). One of these strands is selectively loaded into AGO to form the RNA-induced silencing complex (RISC). Alternatively, AGO2 has been shown to directly cleave several pre-miRNAs, which are then further processed by the poly(A)-specific ribonuclease (PARN) to give rise to mature miRNAs. This class of miRNAs is known as the AGO2-cleaved pre-miRNAs (ac-pre-miRNAs) [78-80, 181]. Such DICER-independent processing involves the pre-miRNA deposit complex (miPDC), which is composed of just AGO2 and a pre-miRNA [77].

The biogenesis of miRNAs is further complicated by the fact that both pre- and mature miRNAs can be post-transcriptionally modified at the 3' end by trimming or by non-templated addition of ribonucleotides, especially uridine and adenine [69, 182-185]. In general, mono-uridylation is thought to re-establish a 3'-2 nucleotide overhang, which is required for efficient DICER cleavage. In contrast oligo-uridylation has been shown to be a signal for degradation and usually occurs after AGO2-mediated slicing of pre-miRNAs. However, most detailed studies of

pre-miRNAs have been performed on a small number of pre-miRNA sequences, thus our understanding of the overall landscape, and global function of these modifications remains elusive.

High-throughput sequencing of total or AGO-bound smRNAs in numerous cells, tissues, and organisms has provided a wealth of information about miRNAs. In concert with bioinformatic approaches, these datasets have been leveraged to identify thousands of novel miRNAs [186, 187]. In contrast, sequencing of pre-miRNAs has been challenging, due to the presence of other RNA species, including the highly abundant transfer RNAs (tRNAs) and small nucleolar RNAs (snoRNAs), that exist in the same size range as pre-miRNAs. Attempts to use size selection to sequence pre-miRNAs have achieved < 1% of total sequenced clones corresponding to pre-miRNAs even after selective depletion of abundant species [188]. Alternatively, primer-based approaches have been applied, but these methods restrict analysis to known miRNA species and cannot be used for discovery [184, 185, 189]. Thus, there is a need for unbiased, sequencing based approaches to gain a more comprehensive understanding of pre-miRNA expression and sequence content.

Leveraging the knowledge that AGO is an integral component of the pre-miRNA processing complexes (miRLC and miPDC) and the miRNA functional (miRISC) complex, our groups recently developed an approach to enrich for pre-miRNAs by immunoprecipitating AGO proteins and isolating RNAs from 50-80 nucleotides (nts) [183]. Using this approach, pre-miRNA libraries from mouse embryonic fibroblasts (MEFs) were generated with > 40% of reads mapping to miRNA loci. Thus, this strategy can be used to efficiently study pre-miRNAs globally without the need for primer-based approaches or depletion of abundant RNA species. Here, we applied this approach to isolate and sequence pre-miRNAs and mature miRNAs from human embryonic kidney (HEK293T) cells. We developed a bioinformatic pipeline to capture post-transcriptional modifications, which we applied to data generated in this study from human cells, as well as to previous data generated in MEFs. Our results provide global insights into pre-miRNA processing

and provide an alternative strategy for identifying pre-miRNAs and other AGO-associated stem-loops in transcriptomes of interest.

4.2 RESULTS

4.2.1 Isolation and sequencing of pre-miRNAs

To isolate and sequence pre-miRNAs we first immunoprecipitated AGO proteins from HEK293T cells with the pan-AGO-2A8 antibody [190] (Figure 4.1A). RNA was purified from AGO immunoprecipitates, dephosphorylated and labeled with P^{32}-γ-ATP. Autoradiography of the RNA gel showed a major band at 20-25 nts, representing mature miRNAs, and several other prominent bands between 50-80 nts, corresponding to the size range of pre-miRNAs (Figure 4.1B). We excised gel slices from both of these regions and generated high-throughput sequencing libraries, referred to herein as miRNA-seq and pre-miRNA-seq libraries, respectively. Previous attempts to sequence pre-miRNAs have been unsuccessful due to the inaccessibility of the pre-miRNA 5' ends. Therefore, we used a method that attaches the 5' linker through CircLigase-mediated cDNA circularization step (see Methods) [13, 183].

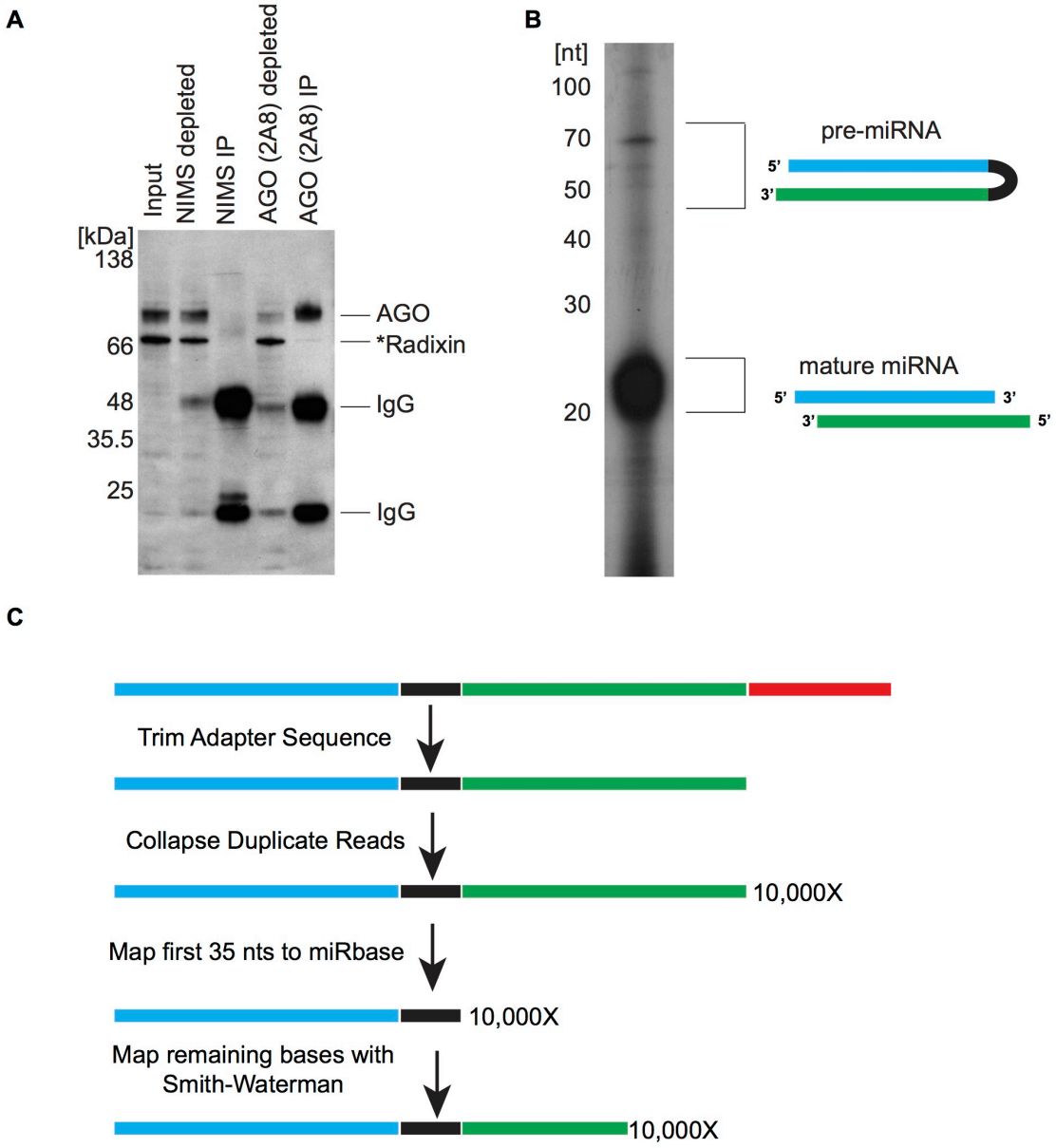


Figure 4.1 Isolation and sequencing of AGO interacting pre-miRNAs and mature miRNAs. A) Western blot of AGO-IP from HEK293T cells. AGO-2A8 and none immune serum (NIMS) were used for immunoprecipitaiton. AGO-2A8 antibody was used for detection. *Radixin is known to cross-react with the 2A8 antibody. B) Autoradiography of RNA co-immunoprecipitated with AGO. Pre-miRNAs were excised from 50-80 nts and mature miRNAs were excised from 20-25 nts. C) Bioinformatics pipeline: Adapter sequences were trimmed and PCR duplicates were collapsed for efficiency. The first 35 nts of pre-miRNAs (18 nts of mature miRNAs) were mapped to miRBase (v20). Remaining bases were mapped with Smith-Waterman aligner.

Using this approach, we successfully generated high-throughput sequencing libraries for both pre-miRNAs and mature miRNAs. We obtained 27.8 and 9.8 million reads with sufficient adapter sequence from the pre-miRNA-seq and mature miRNA-seq libraries, respectively (Table 4.1). Given that pre-miRNAs and mature miRNAs contain non-templated modifications at their 3' ends, we reasoned that standard alignment pipelines would be limited in their ability to align these sequences to miRBase. Therefore, we developed an alignment pipeline that utilized the first 35 nts of the pre-miRNA-seq (18 nts for miRNA-seq) reads for alignment (Figure 4.1C), followed by Smith-Waterman local alignment to extend the read as far as possible along the mapped miRNA sequence (see Methods) [191-193]. After all possible nucleotide matches were made, we selected alignments with the lowest mismatch rate and captured non-templated modifications at the 3' end.

We performed this analysis on pre-miRNA-seq, miRNA-seq, and smRNA-seq (without IP; Vandivier et al. Under Review) from human HEK293T cells as well as pre-miRNA-seq and mature miRNA-seq data previously generated using the same technique from MEFs [183]. We successfully aligned 10.8% of pre-miRNA-seq, 98.7% of miRNA-seq, and 52.4% of smRNA-seq reads to the human miRBase v20 annotation set (Figure 4.2 and Table S1). We were not surprised to find such high rates of mapping for mature miRNAs, however, a 10% mapping rate for pre-miRNAs is significantly higher than previous attempts to sequence pre-miRNAs without pre-miRNA specific primers (0.8%) [188]. Using the datasets previously generated from MEFs, we had a much higher rate of pre-miRNA-seq reads mapping (44%), but mapped fewer miRNA-seq reads (90.3%) (Figure 4.2 and Table S1). These differences likely represent variable experimental conditions and amounts of starting material and/or biological differences in the smRNA populations between these two mammals.

Table 4.1 Summary of sequencing libraries and mapping statistics

Library	Trimmed (reads)	Mapped to miRbase (reads)	Mapped to miRbase (%)	Mapped to RefSeq (reads)	Mapped to RefSeq (%)
HEK293T pre-miRNA-seq	27,836,261	3,009,275	10.81%	1,208,362	4.34%
HEK293T miRNA-seq	9,862,300	9,731,649	98.68%		
HEK293T smRNA-seq	50,887,440	26,643,855	52.36%		
MEF pre-miRNA-seq	16,597,532	7,306,162	44.02%	368,200	2.22%
MEF miRNA-seq	22,044,132	19,906,265	90.30%		

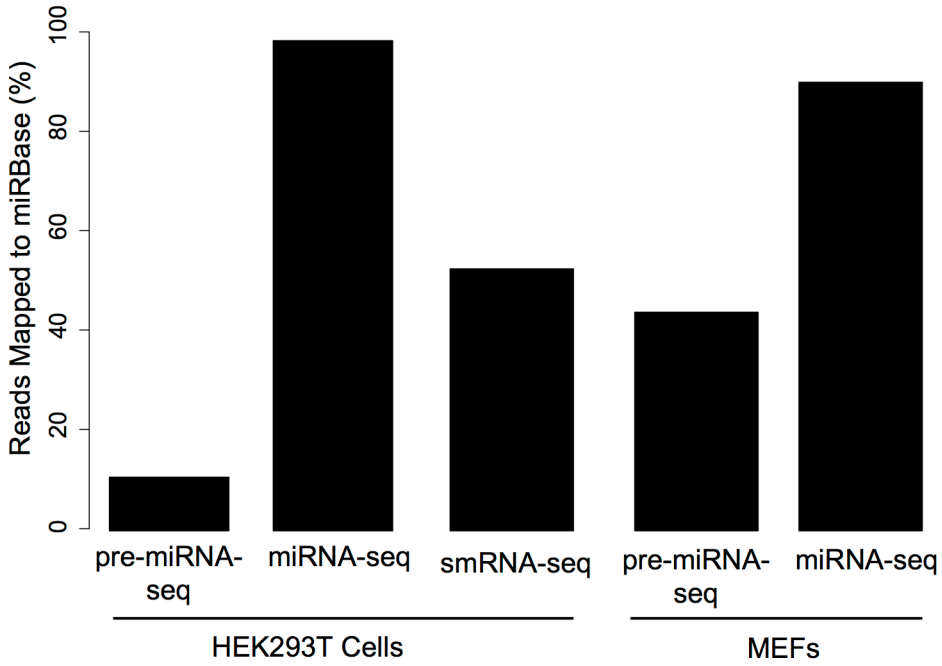


Figure 4.2 Percent of reads mapping to miRBase for pre-miRNA-seq, miRNA-seq and smRNA-seq.

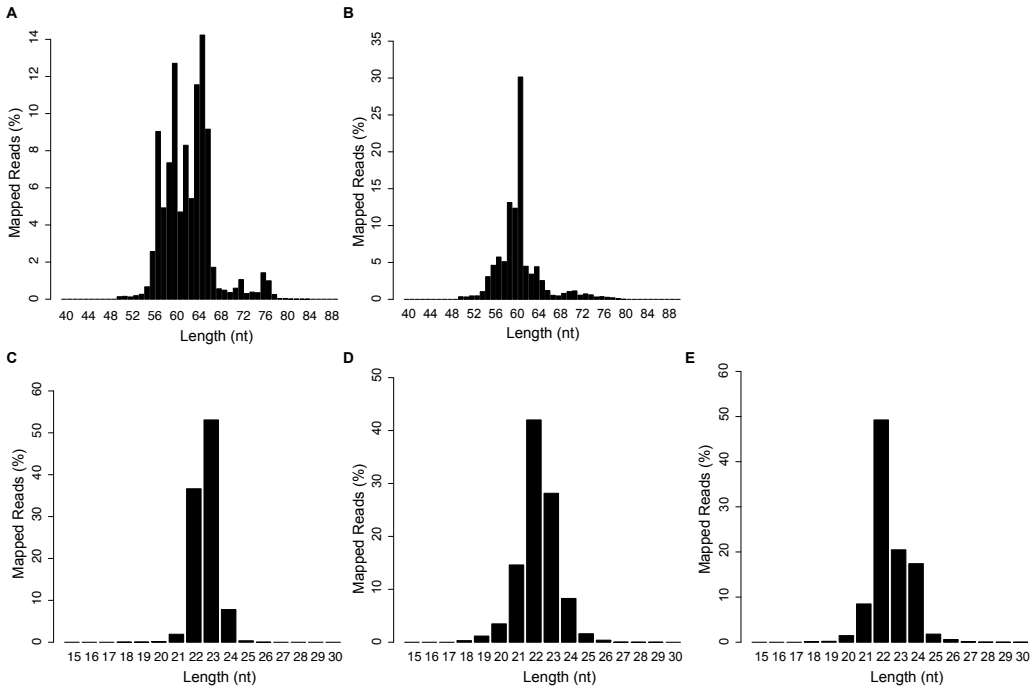


Figure 4.3 Size distribution of miRBase mapped reads. A-B) Size distribution of pre-miRNA-seq reads from HEK293T cells (A) and MEFs (B) mapping to miRBase. C-E) Size distribution of miRNA-seq reads from HEK293T cells (C) and MEF (D) and smRNA-seq reads from HEK293T cell (E).

Overall, we obtained pre-miRNA-seq reads mapping to 367 annotated human miRNAs and 267 annotated mouse miRNAs. For libraries prepared from smaller RNA species, we mapped reads to 931 (HEK239T miRNA-seq), 567 (MEF miRNAs-seq), and 1,364 (HEK293T smRNA-seq) miRBase miRNAs. We examined the size distribution of mapped pre-miRNA-seq reads and found that they were distributed between 55-65 nts in both cell types (Figure 4.3A-B). MiRNA-seq reads, were tightly distributed between 21-24 nts, in both HEK293T AGO-bound and total cellular fractions, as well as in MEFs (Figure 4.3C-E)**.** We determined the abundance, end concordance, and non-templated additions for all mapped miRNAs and generated coverage plots to represent these data, which are available for download at http://gregorylab.bio.upenn.edu/AGO_IP_Seq/ (examples in Figure 4.4). Together, our biochemical and bioinformatic approaches provide a data rich resource for the global and unbiased analysis of pre-miRNAs in two mammals.

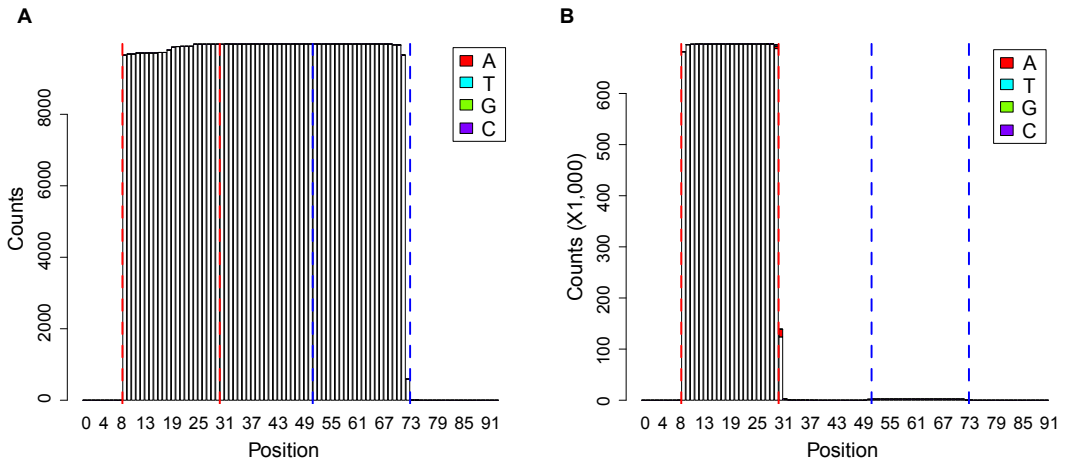


Figure 4.4 Pre-miRNA-seq and miRNA-seq coverage of hsa-miR-16-2. A-B) Coverage plot of pre-miRNA-seq (A) and miRNA-seq (B) reads mapping to hsa-miR-16-2 locus. White bars indicate templated nucleotides. Colored bars indicate non-templated additions. Dashed red and blue lines indicate boundaries of annotated 5p (blue) and 3p (red) mature miRNAs

4.2.2 Diverse ends of AGO-bound pre-miRNAs

It is well established that pre-miRNA trimming and non-templated tailing (uridylation) is a mechanism of regulation [69, 182, 184]. However, most studies have investigated individual miRNAs or used targeted approaches to examine a predetermined subset of miRNAs. Using our novel datasets, we examined the end concordance of sequenced pre-miRNA-seq (Figure 4.5A-B) and miRNA-seq reads (Figure 4.5C-D), relative to high confidence human and mouse miRBase miRNA ends. We found that the majority (>90%) of 5' read ends of pre-miRNAs coincided with annotated 5' ends of 5p miRNAs for both human and mouse (Figure 4.5A-B). In contrast 3' read ends of pre-miRNAs were highly variable, with only 40% and 20% of reads ending precisely at the annotated 3' end of 3p miRNAs in HEK293T cells and MEFs, respectively (Figure 4.5A-B). Relatedly, mature miRNAs (Figure 4.5C-D) and smRNAs (Figure 4.5E) displayed higher variation in their 3' ends relative to 5' ends for both 5p and 3p miRNAs, but not nearly to the extent that was observed for pre-miRNAs (Figure 4.5A-B). Overall, these results reveal that 3' end variation is more common in pre-miRNAs than in mature miRNAs.

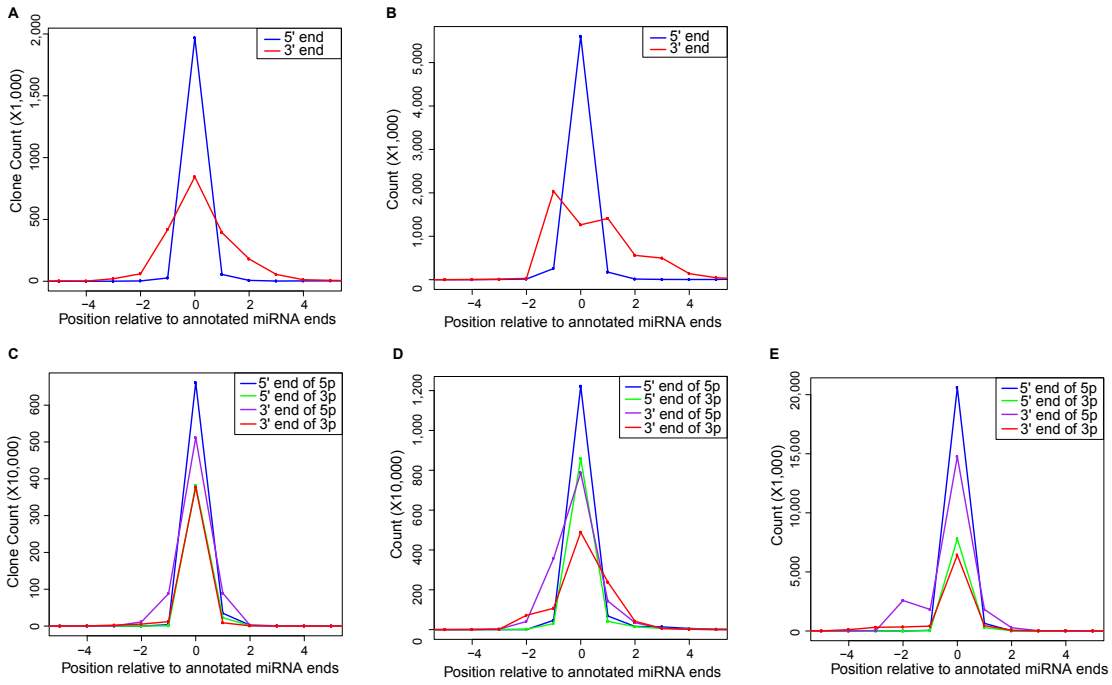


Figure 4.5 Mapped read ends relative to miRBase miRNAs. A-B) Distribution of pre-miRNA-seq read ends for HEK293T cells (A) and MEFs (B) relative to annotated 5' end of 5p and 3' end of 3p miRNAs for high confidence miRNAs. Positive and negative values indicate trimming and extension of reads respectively. C-E) Distribution of miRNA-seq read ends for HEK293T cells (C), MEFs (D) and smRNA-seq read ends from HEK293T cells (E) relative to annotated ends of 5p and 3p miRNAs for high confidence miRNAs. Positive and negative values indicate trimming and extension of reads respectively.

Pre-miRNAs and miRNAs are post-transcriptionally modified at their 3' ends through the action of TUT4 and TUT7 terminal uriydyl transferases (TUTases) [69, 182, 194, 195]. We found that 14.5% of human and 17.4% of mouse pre-miRNAs contained single nucleotide additions to their 3' ends, whereas 4.9% and 7.1% had more than two non-templated additions on their 3' ends in human and mouse high confidence miRBase miRNAs, respectively (Figure 4.6). For mature miRNAs, we also found a large number of single nucleotide additions at the 3' end; 8.4% of human and 12.5% of mouse mature miRNAs. However, tails greater than one nucleotide were much less frequent, with only 1.1% of human and 2.2% of mouse mature miRNAs containing long tails. For cellular smRNA-seq from HEK293T cells we observed a higher overall levels of modification than miRNA-seq with 4.0% and 3.3% of reads having single or multiple non-

templated additions, respectively (Figure 4.6). These data demonstrate that non-templated additions are widespread in both pre-miRNAs and mature miRNAs, and that extended tails are more frequent in pre-miRNAs compared to mature miRNAs.

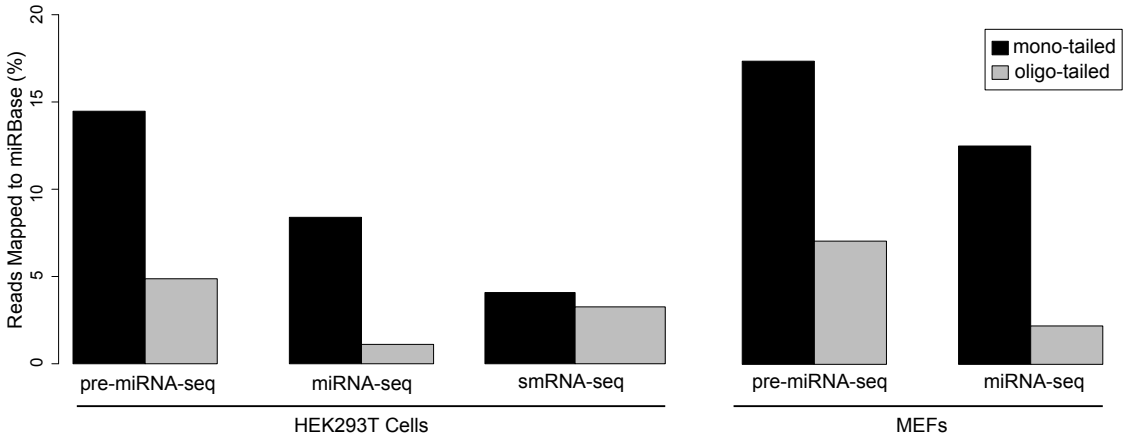


Figure 4.6 Percentage of reads with mono-tails and oligo-tails for pre-miRNA-seq, miRNA-seq and smRNA-seq reads mapped to high confidence miRBase miRNAs.

We also generated metaplots to analyze the sequence content of non-templated additions to the 3' end of pre-miRNAs (Figure 4.7A-B) and mature miRNAs (Figure 4.7C-E). Uridine (denoted as T) was by far the most common addition to human and mouse pre-miRNAs, and was even more prevalent in positions past the first non-templated nucleotide (Figure 4.7A-B). For mature miRNAs, adenosine was the most common non-templated addition, with much lower levels of uridine as compared to pre-miRNAs (Figure 4.7C-E). When examining non-templated additions in total smRNA-seq data, we found that 2 nucleotide tails were much more common than in AGO-interacting miRNAs (Figure 4.7E). Collectively, these data show that mono- and oligo-tailing are widespread in human and mouse pre-miRNAs and mature miRNAs and that uridylation is more common in pre-miRNAs, especially after the first nucleotide, whereas adenylation is more common in mature miRNAs.

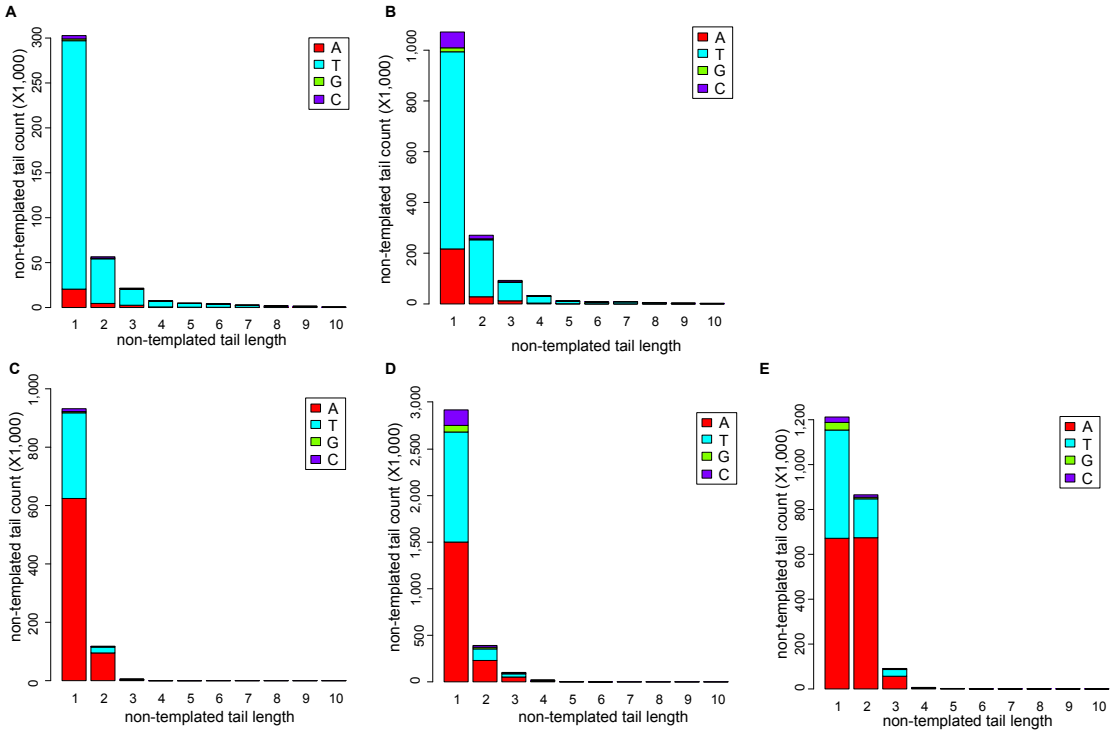


Figure 4.7 Non-templated 3' end modifications. A-B) Non-templated 3' end tail length and sequence content for pre-miRNA-seq reads from HEK293T cells (A) and MEFs (B) reads mapping to miRBase high confidence miRNAs. C-E) Non-templated 3' end tail length and sequence content for miRNA-seq reads from HEK293T cells (C) and MEFs (D) and smRNA-seq reads from HEK293T cells (E) mapping to miRBase high confidence miRNAs.

4.2.3 Identification of AGO2-cleaved pre-miRNAs

AGO2, is unique amongst the AGO proteins in that it has slicing activity [196-198]. In fact, AGO2 has been demonstrated to cleave a subset of pre-miRNAs (ac-pre-miRNAs), which are then trimmed by PARN into mature miRNAs [78-80, 181]. However, the extent to which this process occurs in mammalian pre-miRNA populations has not been determined. AGO2 cleavage events are known to occur around 10 nucleotide upstream of the 3' end of the pre-miRNA and give rise to 5p miRNAs [78]. To search for these events, we calculated the percentage of pre-miRNA-seq reads that were trimmed 8-15 nucleotide from the 3' end of the 3p miRNA for each miRNA. In mouse, for which ac-pre-miRNAs are best characterized, we observed all previously identified ac-pre-miRNAs with the exception of miR-451, which is not expressed in MEFs (Table

4.2). In our human dataset, we also observed previously identified ac-pre-miRNAs, including 9 members of the let-7 family and miR-9-2. From this approach, we identified 7 putative ac-pre-miRNA candidates in humans and 37 in mouse, including hsa-miR-455 and mmu-miR-335 (Figure 4.8A-B and Tables 4.2**).**

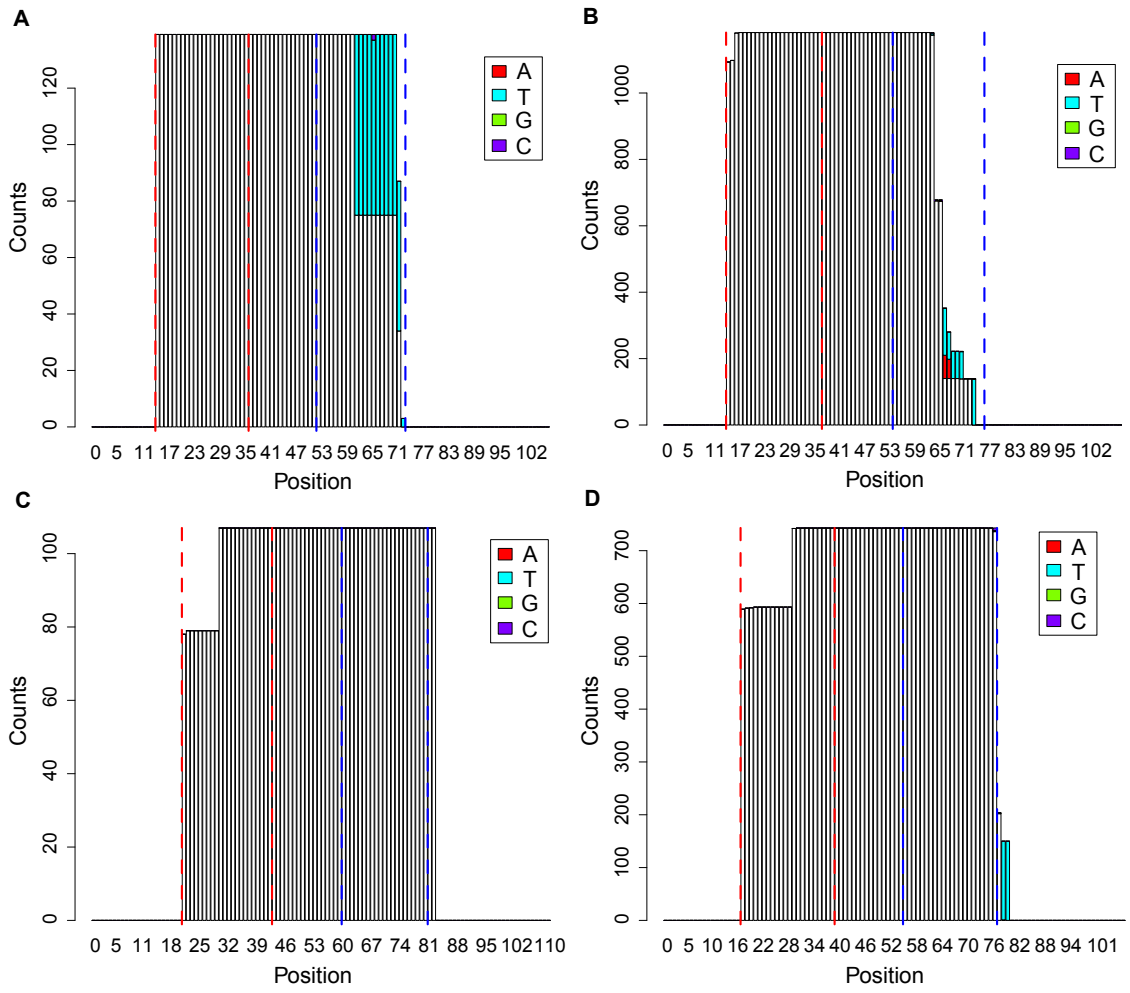


Figure 4.8 De novo identification of AGO2-cleaved pre-miRNAs. A-D) Examples of novel ac-pre-miRNAs identified from pre-miRNA-seq. hsa-miR-455 (A) and mmu-miR-335 (B) are novel ac-pre-miRNAs with cleavage in the 3p miRNA. hsa-miR-140 (C) and mmu-miR-22 (D) are novel ac-pre-miRNAs with cleavage in the 5p miRNA. White bars indicate templated nucleotides, colored bars indicate non-templated additions. Dashed red and blue lines indicate boundaries of annotated 5p (blue) and 3p (red) miRNAs.

Table 4.2 De novo identification of ac-pre-miRNAs in MEFs. Bolded rows indicate known ac-pre-miRNAs

Name	Cleaved Clones	Total Clones	Percent Cleaved	5p miRNAs	3p miRNA	Percent 5p
mmu-let-7a-1	**4943**	**78608**	**6.29%**	**227064**	**27133**	**89.33%**
mmu-let-7b	**37000**	**68610**	**53.93%**	**139401**	**5988**	**95.88%**
mmu-let-7c-1	241	5289	4.56%	296030	23	99.99%
mmu-let-7c-2	**6913**	**173344**	**3.99%**	**166758**	**26178**	**86.43%**
mmu-let-7d	**2665**	**12663**	**21.05%**	**364660**	**63447**	**85.18%**
mmu-let-7e	242	1694	14.29%	20622	2712	88.38%
mmu-let-7f-1	**25368**	**56223**	**45.12%**	**129389**	**4943**	**96.32%**
mmu-let-7i	**1264**	**1795**	**70.42%**	**125465**	**11219**	**91.79%**
mmu-mir-101a	**558**	**879**	**63.48%**	**638**	**5105**	**11.11%**
mmu-mir-106b	7784	100647	7.73%	160155	9516	94.39%
mmu-mir-10a	1134	34484	3.29%	85754	2319	97.37%
mmu-mir-125b-1	389	13989	2.78%	939092	45080	95.42%
mmu-mir-125b-2	2729	4277	63.81%	933810	2376	99.75%
mmu-mir-181a-2	627	35394	1.77%	62245	4109	93.81%
mmu-mir-181b-1	1342	51785	2.59%	23557	83	99.65%
mmu-mir-181b-2	272	6842	3.98%	28160	99	99.65%
mmu-mir-183	323	3156	10.23%	16348	220	98.67%
mmu-mir-1839	169	509	33.20%	3225	897	78.24%
mmu-mir-188	139	4641	3.00%	4085	98	97.66%
mmu-mir-18a	274	18156	1.51%	79317	4586	94.53%
mmu-mir-195a	147	2585	5.69%	1224	736	62.45%
mmu-mir-199a-2	271	25099	1.08%	148775	873207	14.56%
mmu-mir-26a-2	5060	124040	4.08%	238340	651	99.73%
mmu-mir-28a	25	703	3.56%	32209	3247	90.84%
mmu-mir-297a-4	184	185	99.46%	7607	3107	71.00%
mmu-mir-297c	168	169	99.41%	2412	3107	43.70%
mmu-mir-3079	25	110	22.73%	31	0	100.00%
mmu-mir-30e	62	3749	1.65%	69001	1428	97.97%
mmu-mir-31	**2161**	**40063**	**5.39%**	**449467**	**69169**	**86.66%**
mmu-mir-322	**835**	**1561**	**53.49%**	**22178**	**13211**	**62.67%**
mmu-mir-335	1006	1141	88.17%	627	309	66.99%
mmu-mir-345	**229**	**1798**	**12.74%**	**16236**	**434**	**97.40%**
mmu-mir-374b	73	1125	6.49%	70335	423	99.40%
mmu-mir-449c	160	530	30.19%	100	25	80.00%
mmu-mir-452	219	549	39.89%	465	312	59.85%
mmu-mir-466c-1	159	312	50.96%	1424	4708	23.22%
mmu-mir-467b	1113	2775	40.11%	6330	45	99.29%
mmu-mir-467c	186	1644	11.31%	2763	342	88.99%
mmu-mir-467d	145	3557	4.08%	3200	2352	57.64%
mmu-mir-467e	184	4482	4.11%	4354	380	91.97%
mmu-mir-503	1816	71861	2.53%	13223	496	96.38%
mmu-mir-542	197	586	33.62%	271	623	30.31%
mmu-mir-669c	106	645	16.43%	336	50	87.05%
mmu-mir-669d-2	9	514	1.75%	4327	3	99.93%
mmu-mir-669h	54	262	20.61%	68	161	29.69%
mmu-mir-669p-1	5	122	4.10%	6091	325	94.93%
mmu-mir-7a-1	103	8399	1.23%	7386	3052	70.76%
mmu-mir-9-2	**5453**	**9481**	**57.52%**	**29673**	**3426**	**89.65%**
mmu-mir-98	**877**	**964**	**90.98%**	**7122**	**2399**	**74.80%**

All previous examples of ac-pre-miRNAs are cleaved in the 3p miRNA and generate 5p miRNAs. To identify putative ac-pre-miRNAs that are processed in the opposite directionality, we performed a parallel search for pre-miRNA-seq reads that were trimmed 8-15 nucleotide from the 5' end of the 5p miRNA. From this analysis, we found 3 candidates in humans and 5 candidates in mouse including hsa-miR-140 and mmu-miR-22 (Figure 4.8C-D). Importantly, these 5p cleaved ac-pre-miRNAs give rise predominantly to 3p miRNAs suggesting that cleavage in this region is not a degradation byproduct and is likely a competent mechanism for generating mature miRNAs. These results reveal a greater collection of ac-pre-miRNAs than previously appreciated and uncover a likely novel class of 5p cleaved ac-pre-miRNAs.

4.2.4 Relating pre-miRNA and mature miRNA abundance

Almost nothing is known about the relationship between pre-miRNA and mature miRNA abundance. In order to assess this, we first grouped miRNAs into their families and merged the mapped reads. This was necessary to avoid artifacts, given that multiple distinct precursors give rise to identical, or nearly identical mature miRNAs. We further refined our analysis by only focusing on high confidence miRNAs, which left us with 158 human and 159 mouse miRNAs with reads mapping from pre-miRNA-seq and/or miRNA-seq libraries. We analyzed the correlation between pre-miRNAs and mature miRNA levels expressed from these families (Figure 4.9A-B). We found a positive but modest correlation between pre-miRNA and mature miRNA expression in both humans (Spearman correlation R = 0.53, p-value < $6.635e^{-08}$) and mouse (Spearman correlation R = 0.52, p-value < $2.047e^{-07}$). We also assessed the relationship between smRNA-seq and pre-miRNA-seq in HEK293T cells (Figure 4.9C). We found a strikingly similar correlation to that of miRNA-seq (Spearman correlation; R = 0.53, p-value < $4.331e^{-08}$), which is explained by the high correlation of smRNA-seq to AGO-IP-seq (Spearman correlation; R = 0.90, p-value < $2.2e^{-16}$) (Figure 4.9D). Together, these results reveal that pre-miRNA and mature miRNA levels have a modest positive correlation in both mouse and humans.

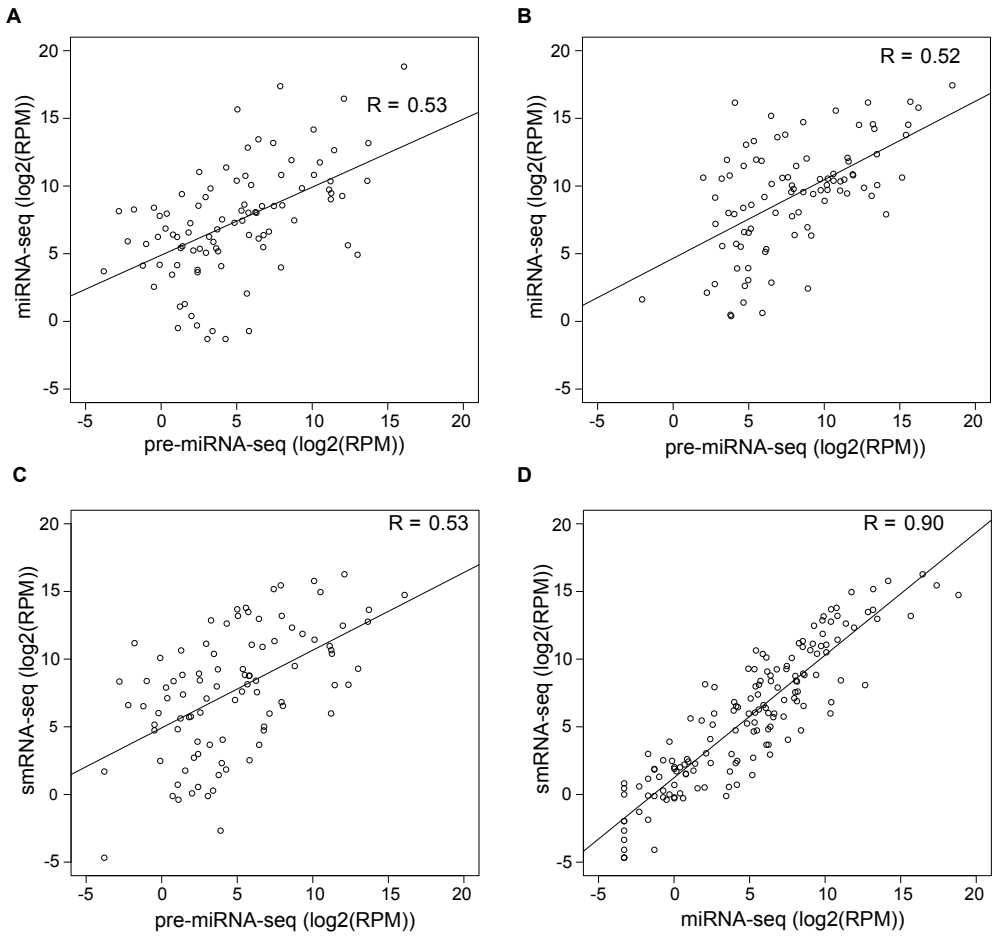


Figure 4.9 Relating pre-miRNA and miRNA abundance. A-B) Correlation of pre-miRNA-seq and miRNA-seq for high confidence human (A) and mouse (B) miRNA loci (human: R = 0.53, p-value < $6.635e^{-08}$, mouse: R = 0.52, p-value < $2.047e^{-07}$; Spearman correlation coefficient). C) Correlation of pre-miRNA-seq and smRNA-seq for high confidence human miRNA loci (Spearman correlation R = 0.53, p-value < $4.331e^{-08}$). D) Correlation of miRNA-seq and smRNA-seq for high confidence human human miRNA loci (Spearman correlation R = 0.90, p-value < $2.2e^{-16}$).

4.2.5 An index for miRNA precursor processing

We reasoned that the ratio of mature miRNAs to pre-miRNAs represents a reasonable estimate for *in vivo* miRNA processing efficiency. As some miRNAs had no detectable pre-miRNA reads, we used a generalized log odds ratio of miRNA-seq to pre-miRNA-seq reads (see METHODS) to compute a miRNA precursor processing index (MPPI) for miRNAs expressed in

human HEK293T and mouse MEF cells. As expected, the majority of high confidence miRNAs from human and mouse cells exhibited MPPI values > 0 (Figure 4.10A-C), suggesting they are efficiently processed. Among the maximum scoring high confidence miRNAs in humans was hsa-miR-338, which had no detectable pre-miRNA-seq reads and 1,358 reads per million (RPM) in miRNA-seq libraries (MPPI = 11.4). In contrast, we found a number of miRNAs that had many more pre-miRNA-seq reads than mature miRNA-seq reads. For example, the least efficiently processed high confidence miRNA in humans was hsa-miR-1296, for which we obtained 8,189 RPM in pre-miRNA-seq libraries and only 30 RPM for miRNA-seq libraries, (MPPI = -8.1). For mouse, we found a similar range of MPPI scores with mmu-miR-214 being the most efficiently processed (MPPI = 16.1) and mmu-miR-3572 being the least efficiently processed (MPPI = -9.1). Thus, examining the ratio of mature to pre-miRNAs allows us to determine the efficiency of miRNA processing for hundreds of miRNAs at once.

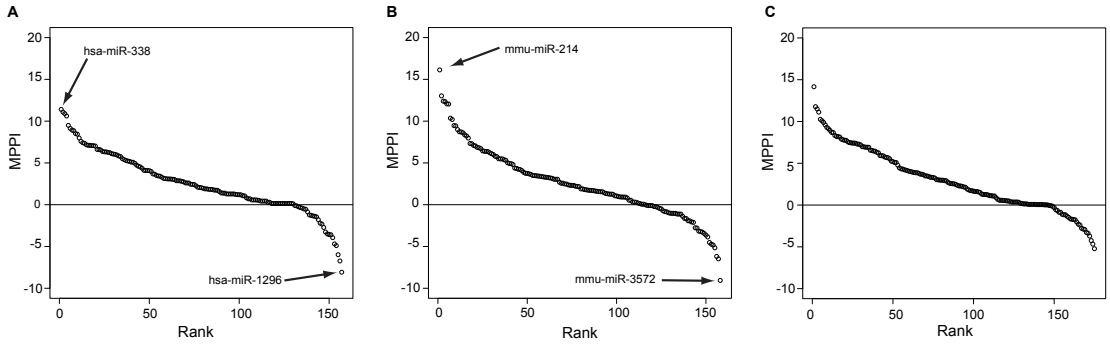


Figure 4.10 An index for miRNA precursor processing efficiency. A-B) Rank ordered list of miRNA precursor processing index (MPPI) scores for high confidence human (A) and mouse (B) miRNAs. C) Rank ordered list of MPPI for high confidence human miRNAs with smRNA-seq.

4.2.6 Distinct processing of two pre-miRNAs from Dgcr8 mRNA

The microprocessor complex (DGCR8/DROSHA) auto-regulates the expression of the *Dgcr8* transcript by binding and cleaving two hairpins near the 5' end of this mRNA [35, 199, 200]. We re-examined the abundance, modification status, and MPPI scores for the two annotated

miRNAs in this region (Figure 4.11). We determined that hsa-miR-3618, which is encoded in the 5' untranslated region (UTR) of *Dgcr8*, is inefficiently processed into mature miRNAs (MPPI = -5.22), whereas hsa-miR-1306, which lies in the coding sequence (CDS), was matured efficiently (MPPI = 3.09) (Figure 4.11). Furthermore, we found that less than 10% of hsa-miR-3618 pre-miRNAs contained non-templated tails, whereas 100% of the hsa-miR-1306 clones were mono-uridylated, a signal that has previously been linked to efficient processing in some miRNAs [69]. Thus, divergent processing of two miRNAs from the same primary transcript underscores the selectivity of DICER processing, and suggests that these two hairpins likely have evolved distinct functions in post-transcriptional regulation.

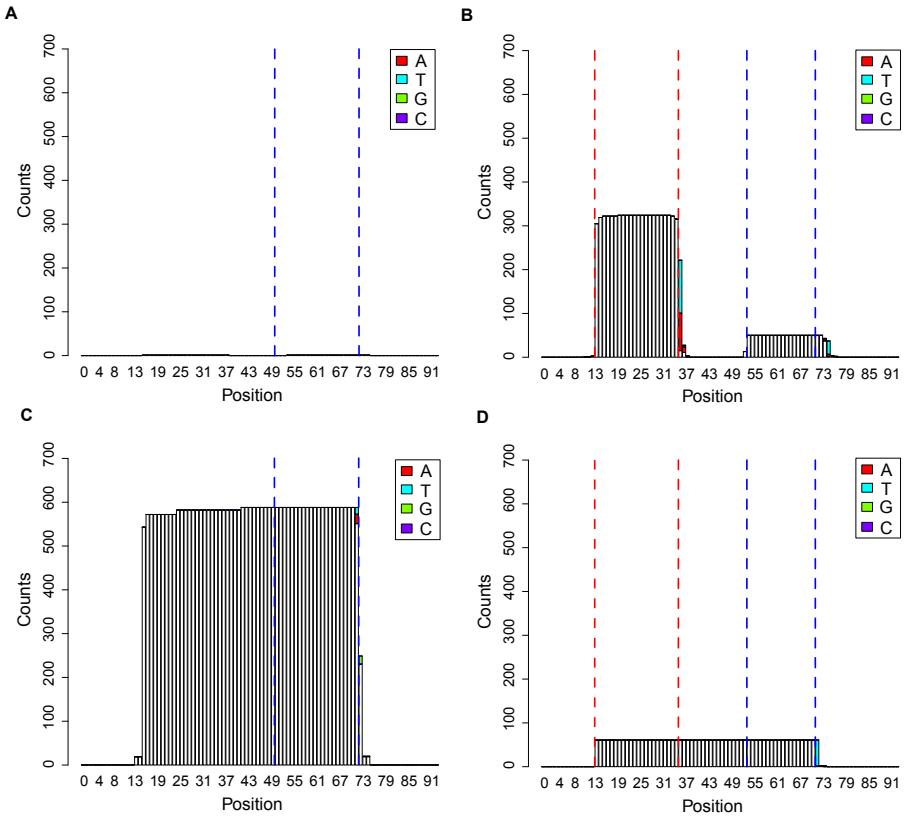


Figure 4.11 Two miRNAs embedded in the Dgcr8 mRNA are processed with highly divergent efficiencies. A-B) miRNA-seq reads mapping to hsa-miR-3618 in the 5' UTR (A) or hsa-miR-1306 in the CDS (B) of *Dgcr8* mRNA. C-D) pre-miRNA-seq reads mapping to hsa-miR-3618 in the 5' UTR (C) or hsa-miR-1306 in the CDS (D) of *Dgcr8* mRNA.

4.2.7 Identification of AGO-associated stem-loops in mRNAs

Given this observation, we examined our dataset for novel pre-miRNAs embedded in other mRNAs. To do this, we took a highly conservative approach, using pre-miRNA-seq reads that failed to map to miRBase, and filtering them by mapping to small nuclear RNAs (snRNAs), snoRNAs, tRNAs, ribosomal RNAs (rRNAs), and repeat-masked sequences (Figure 4.12**,** see Methods). Approximately 1.2 and 0.37 million clones passed our stringent filtering steps and mapped to human and mouse mRNAs, respectively (Table 4.1). To identify significant AGO-associated stem-loops in mRNAs, we used a CLIP-seq peak calling approach to identify significant (modified false discovery rate (mFDR) < 0.01) read clusters in mRNAs [151]. We found a number of highly significant peaks that corresponded to nearly the full length of some highly expressed genes (e.g Actb). Therefore, we further filtered significant clusters based on their size (< 200 nt), then chose the top clone from each cluster, folded it using RNAfold, and captured clusters that had a minimum free energy (MFE) < -0.3 kcal/mol/nt and a minimum of 15 base-pairs in the longest hairpin [201].

This resulted in 403 AGO-associated stem-loops in human and 373 in mouse. We intersected our list of human AGO-associated stem-loops with recently identified miRNAs in humans from smRNA-seq and miRNA prediction or from DICER PAR-CLIP [187, 202]. In fact, 12 of our AGO-associated stem-loops were annotated as novel miRNAs in these lists supporting the validity of our approach (Figure 4.12). Furthermore, 34 and 37 AGO-associated stem-loops overlapped with DICER and DGCR8 binding sites respectively, revealing that a number of the AGO-associated stem-loops interact with other components of the canonical miRNA processing pathway [89, 202].

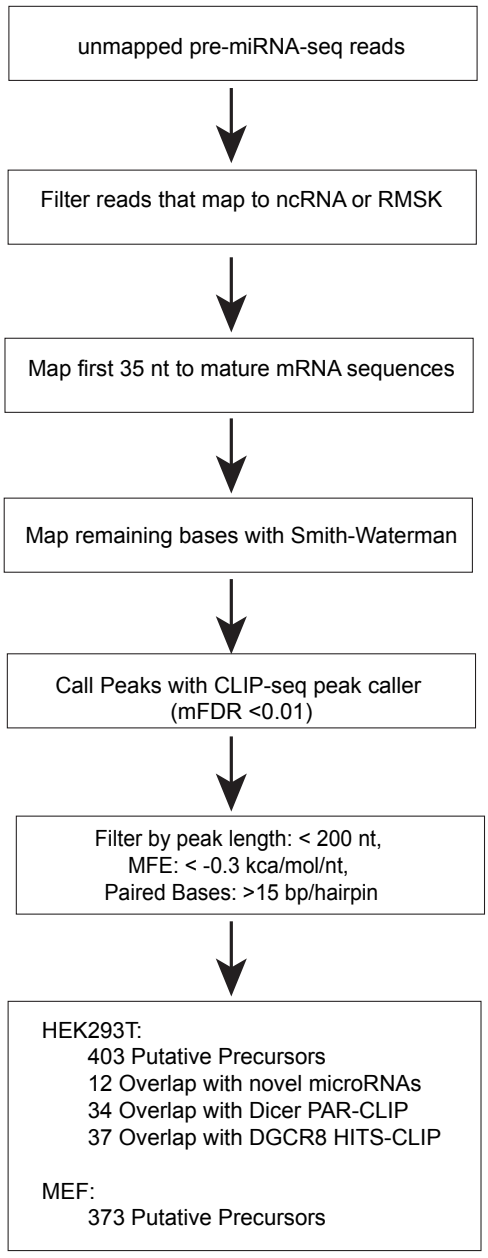


Figure 4.12 Bioinformatics pipeline for identification of AGO-associated stem-loops. Pre-miRNA-seq reads that did not map to miRBase were filtered on ncRNA and repeat-masker (RMSK), trimmed to 35 nts and mapped to spliced mRNA sequences. These alignments were then extended to find all matching bases. AGO-associated stem-loops were identified using a CLIP-seq peak caller (mFDR < 0.01) followed by filtering by length (< 200 nt), MFE (< -0.3 kcal/mol/nt), and paired bases (> 15 bp/hairpin).

We next examined the distribution of AGO-associated stem-loops across mRNAs and found that they were equally present in all regions of mRNAs and similarly distributed between human and mouse (Figure 4.13A). When we normalized the distribution of AGO-associated stem-loops by relative genomic coverage of each mRNA region and found that the CDS was underrepresented, whereas the 5' UTR was enriched 2.5 fold for AGO-associated stem-loops in both organisms (Figure 4.13B). We also examined the size distribution of reads mapping to AGO-associated stem-loops and found them to be similar in size in both mammals, between 55-75 nt in humans and 52-80 nt in mouse (Figure 4.13C-D). This size range was slightly broader than pre-miRNA-seq reads mapped to miRBase (Figure 4.3A-B).

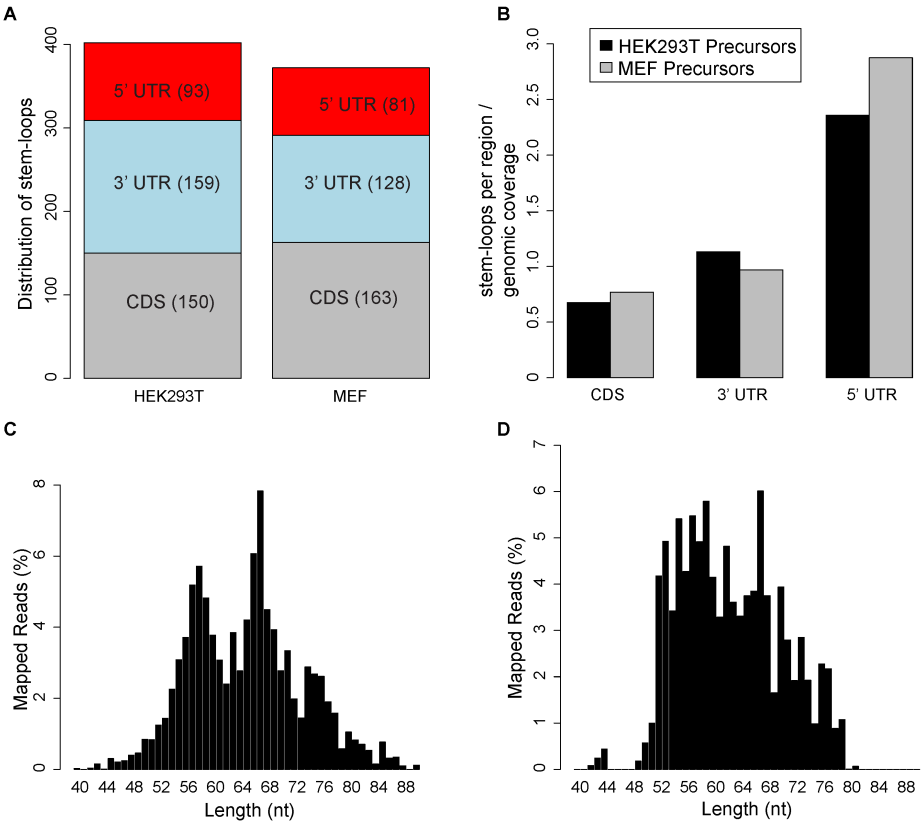


Figure 4.13 Characterization of AGO-associated stem-loops. A) Distribution of AGO-associated stem-loops in mRNAs. B) Enrichment of AGO-associated stem-loops in mRNA regions relative to genomic coverage of mRNA regions. D-E) Size distribution of pre-miRNA-seq reads from HEK293T (D) and MEFs (E) mapping to AGO-associated stem-loops in mRNAs.

We also analyzed the non-templated 3' additions to pre-miRNA-seq reads mapped to AGO-associated stem-loops in human and mouse mRNAs (Figure 4.14A-B). We found that pre-miRNA-seq reads which mapped to AGO-associated mRNA stem-loops were enriched for uridylation events, and in fact a higher percentage were oligo-tailed (13.0% in human and 24.3% in mouse), compared to reads mapping to miRBase (Figure 4.6). This result suggests that the AGO-associated stem-loops in mRNAs identified by our approach undergo similar modifications as known pre-miRNAs.

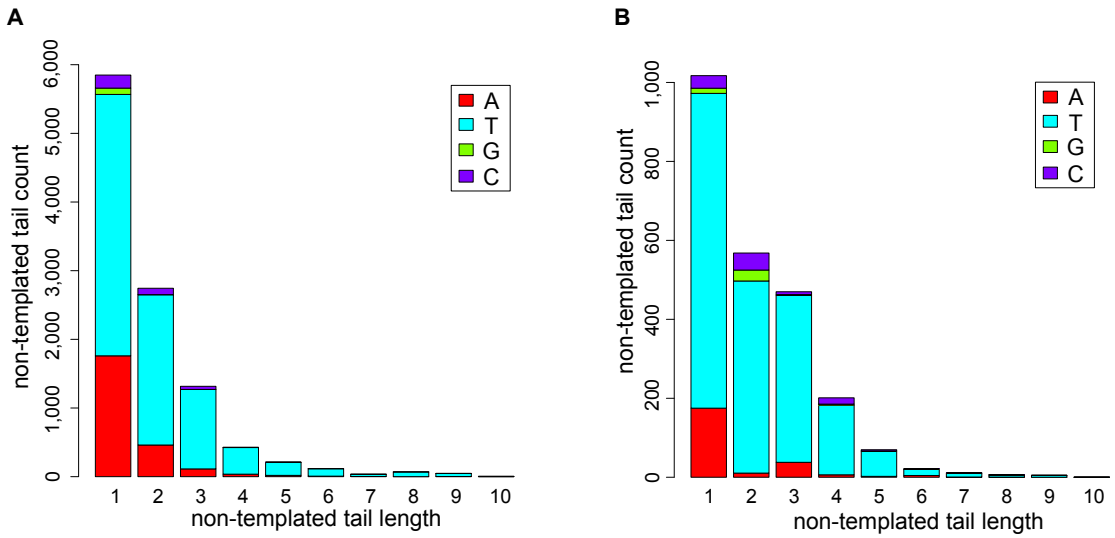


Figure 4.14 Non-templated 3' ends of AGO-associated stem-loops in mRNAs. A-B) Non-templated 3' end tail length and sequence content for pre-miRNA-seq reads mapping to human (A) and mouse (B) AGO-associated stem-loops within mRNAs.

Finally, we overlaid *in vivo* RNA structure-probing data from mouse embryonic stem (ES) cells generated with *in vivo* click selective 2'-hydroxyl acylation and profiling (icSHAPE) onto our AGO-associated stem-loops in mouse mRNAs (Figure 4.15). icSHAPE, chemically modifies the backbone of unpaired nts, causing early termination of reverse transcription [203]. Based on our RNAfold predictions, we examined the icSHAPE reactivity at paired or unpaired bases and found that icSHAPE reactivity was much higher at unpaired bases (Kruskal-Wallis one-way ANOVA; p-value < $8.18e^{-133}$) (Figure 4.15A). This can been seen clearly when examining the RNAfold

structure diagrams of AGO-associated stem-loops with icSHAPE reactivity overlaid (Figure 4.15B-D). Together, our analyses provide numerous candidate pre-miRNA-like elements that are processed into 50-80 nt AGO-associated stem-loops from mRNAs in mammalian cells.

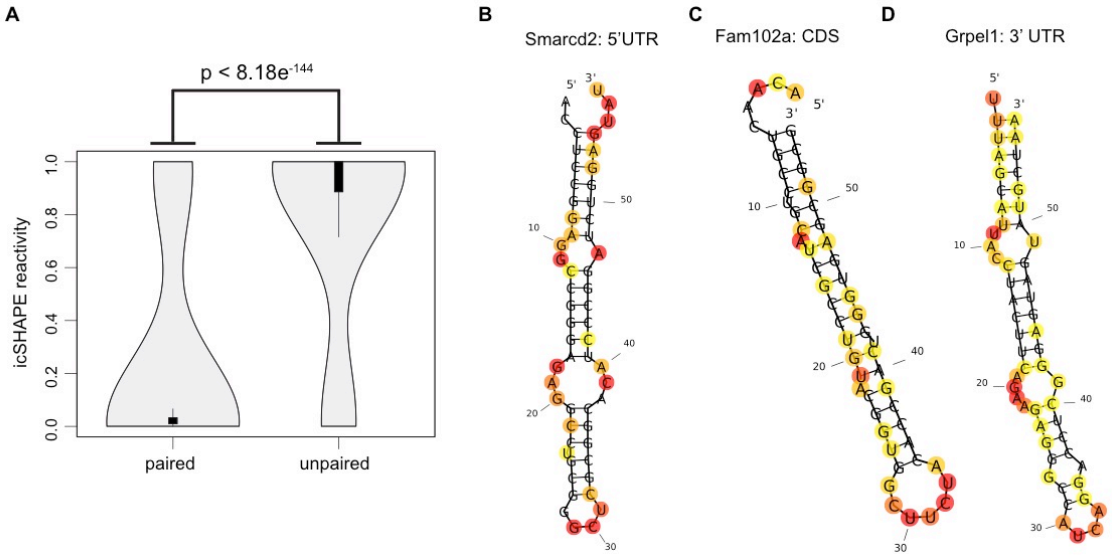


Figure 4.15 icSHAPE supports RNAfold structures of AGO-associated stem-loops in mRNAs. A) Violin plot of *in vivo* icSHAPE reactivity (top and bottom 10^{th} percentile) for paired and unpaired positions in mouse AGO-associated stem-loops in mRNAs. Significance was assessed with Kruskal-Wallis one-way ANOVA; p-value < $8.18e^{-133}$. B-D) Examples of RNAfold predicted structures overlaid with icSHAPE reactivity for AGO-associated stem-loops in the 5' UTR of *Smarcd2* (B), CDS of *Fam102a* (C) and 3' UTR of *Grepl1* (D).

4.2.8 Stem-loop containing mRNAs are regulated by DGCR8 and DROSHA

To assess whether AGO-associated stem-loop containing mRNAs are regulated by the microprocessor complex, we performed siRNA-mediated knockdown of *DROSHA* and *DGCR8* in HEK293T cells and assessed changes in gene expression (Figure 4.16). We observed robust knockdown of both DROSHA and DGCR8 protein levels (Figure 4.16A). Consistent with the role of DROSHA in regulating DGCR8 expression, we found that knockdown of *DROSHA* increased DGCR8 protein and RNA expression (Figure 4.16A-B). Furthermore, RNA-seq data were consistent with RT-qPCR-based validation for both *DROSHA* (Pearson correlation R = 0.99, p-value < 0.0097) and *DGCR8* knockdown (Pearson correlation R = .87, p-value < 0.06) (Figure 4.16B). We found that AGO-associated stem-loop containing mRNAs were significantly

upregulated after knockdown of either *DROSHA* (Wilcoxon rank-sum test, p-value < $1.55e^{-5}$) or *DGCR8* (Wilcoxon rank-sum test, p-value < 0.00047) relative to all genes (Figure 4.16C).

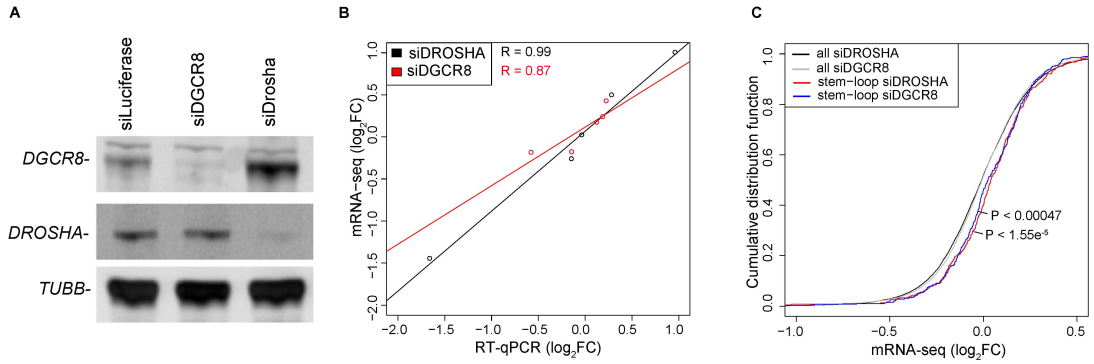


Figure 4.16 mRNAs that host AGO-associated stem-loops are regulated by the microprocessor complex. A) Western blot following siRNA knockdown of *DGCR8*, *DROSHA* and *TUBB* in HEK293T cells. B) Correlation of RT-qPCR and RNA-seq $\log_2$ Fold Change ($\log_2$FC) for selected genes. siDROSHA; Pearson correlation R = 0.99, p-value < 0.0097 and siDGCR8; Pearson correlation R = .87, p-value < 0.06. C) Cumulative distribution function plot of *DROSHA* and *DGCR8* knockdown mRNA-seq $\log_2$FC for AGO-associated stem-loop containing mRNAs compared to all mRNAs. siDROSHA; Wilcoxon rank-sum test, p-value < $1.55e^{-5}$, and siDGCR8; Wilcoxon rank-sum test, p-value < 0.00047.

4.2.9 Most AGO-associated stem-loops in mRNAs do not produce AGO-bound smRNAs

We next calculated the miRNA-seq and smRNA-seq coverage at AGO-associated stem-loops using a similar mapping pipeline as described above for miRNAs. We found that miRNA-seq and smRNA-seq reads mapping to AGO-associated stem-loops were of a similar size as miRBase miRNAs and had non-templated additions to their 3' ends (Figure 4.17). We calculated the MPPI for AGO-associated stem-loops and found them to be inefficient producers of mature AGO-bound smRNAs (Figure 4.18). This is in agreement with the finding that these elements are more commonly oligo-tailed compared to canonical pre-miRNAs. In fact, we only found a total of 171 miRNA-seq reads mapping to AGO-associated mRNA stem-loops in humans and the majority of these reads mapped to the recently identified miRNAs embedded within the mRNAs; *GNAS, GLUL*, and *E2F1* [187, 202] (example of E2F1 in Figure 4.19A-B). Additionally, we found a total of 2,240 smRNA-seq reads mapping to AGO-associated mRNA stem-loops, which in part

reflects the higher sequencing depth of these libraries. Some of putative pre-miRNAs with the most smRNA-seq reads, corresponded to previously identified novel miRNAs, including *BRD2*, *GLUL* and *E2F1* (example in Figure 4.19C). However, we also noticed numerous reads mapping to putative pre-miRNAs that did not have binding evidence from miRNA-seq, including *FTH1*, *FTL*, *SOX4*, and *KMT2C*.

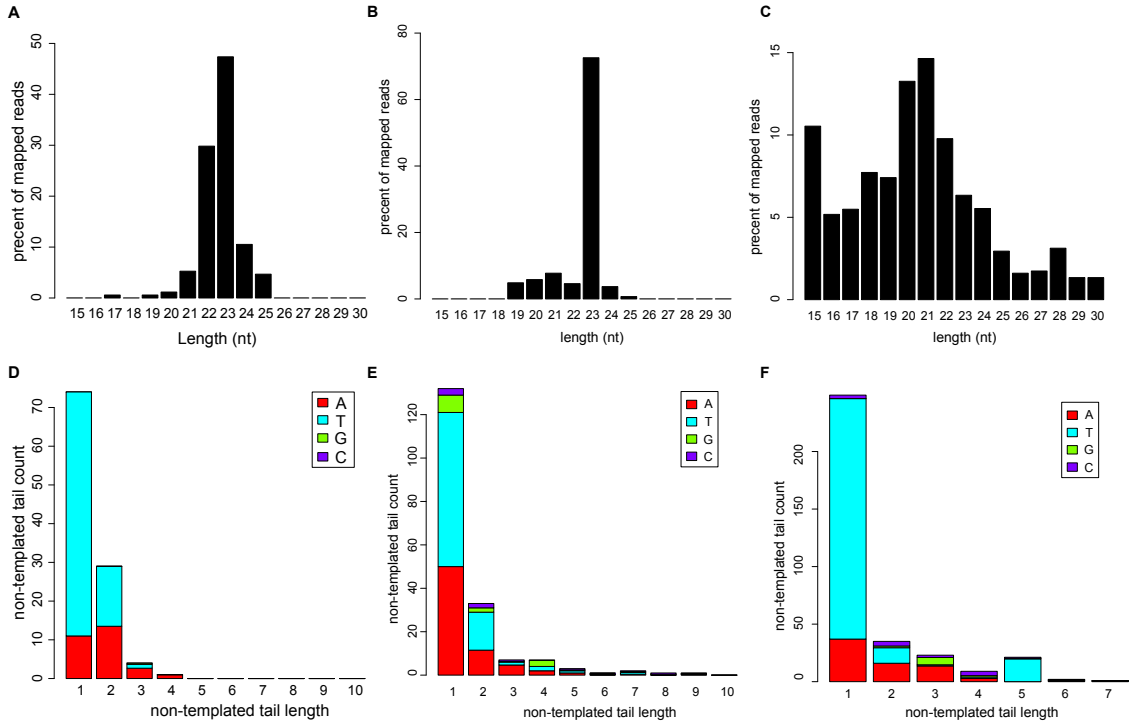


Figure 4.17 Few AGO-associated stem-loops in mRNAs produce smRNAs. A-B) Size distribution of miRNA-seq reads mapping to human (A) and mouse (B) AGO-associated stem-loops in mRNAs. C) Same as in (A) except human smRNA-seq. D-E) Non-templated 3' end tail length and sequence content for miRNA-seq reads mapping to human (D) and mouse (E) AGO-associated stem-loops within mRNAs. F) Same as in (D) except for human smRNA-seq

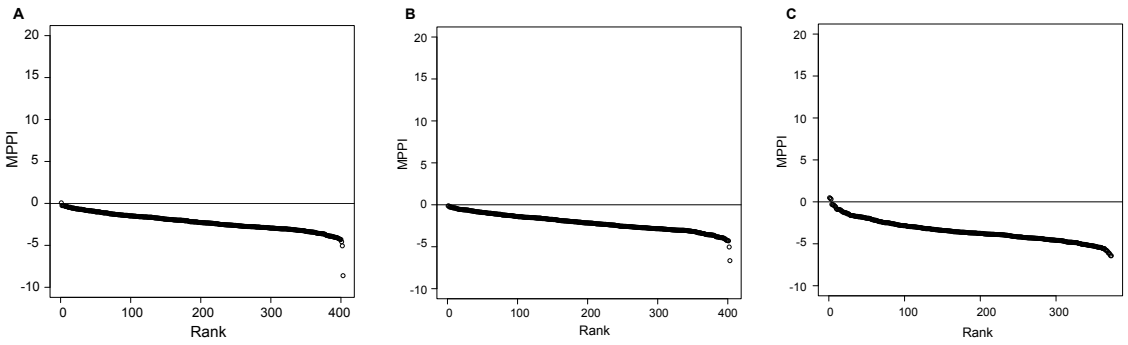


Figure 4.18 AGO-associated stem-loops are inefficient producers of smRNAs. A-B) Rank order of MPPI score for AGO-associated stem-loops from human (A) and mouse (B) mRNAs. Same as in (A) except with human smRNA-seq

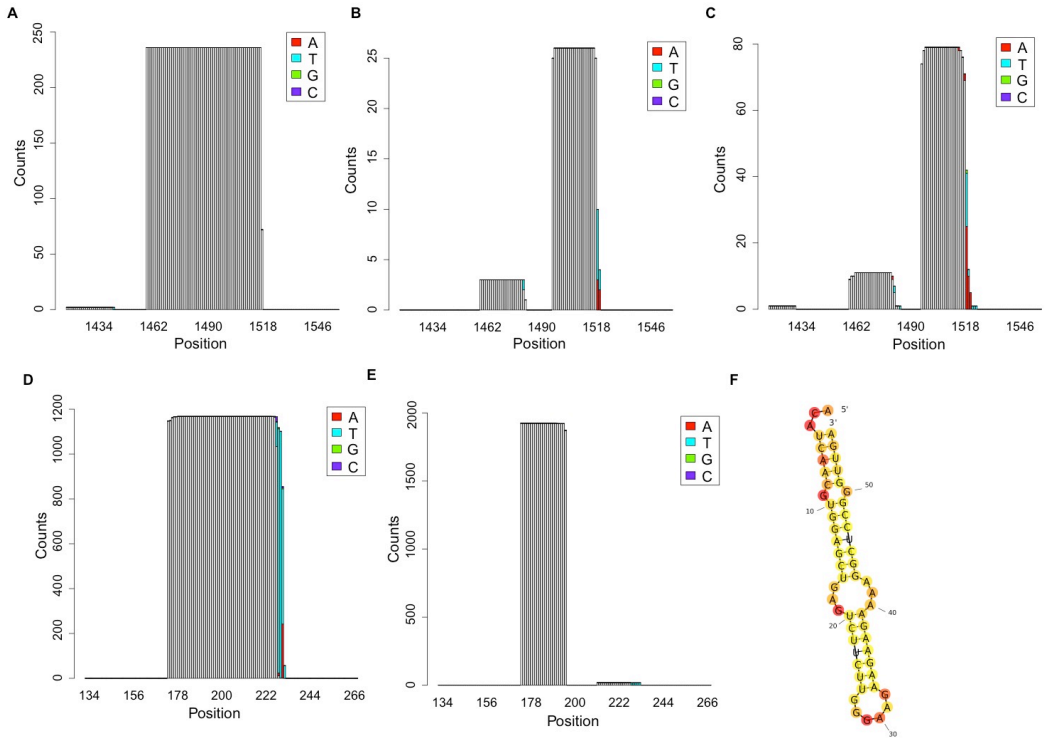


Figure 4.19 pre-miRNA-seq identifies known and novel miRNAs. A-C) Coverage plot of pre-miRNA-seq (A), miRNA-seq (B) and smRNA-seq (C) reads mapping to a human AGO-associated stem-loop in the 3' UTR of *E2F1*. White bars indicate templated nucleotides, colored bars indicate non-templated additions. D-E) Coverage plot of pre-miRNA-seq (D) and miRNA-seq (E) reads mapping to a mouse AGO-associated stem-loop in the CDS of *Rpl9*. White bars indicate templated nucleotides, colored bars indicate non-templated additions. F) RNAfold predicted structure and icSHAPE reactivity for mouse AGO-associated stem-loop in the CDS of *Rpl9*.

For mouse AGO-associated stem-loops we mapped 2,698 miRNA-seq reads, however 71% of these mapped to a single locus in the CDS of *Rpl9*, which likely represents a novel miRNA (Figure 4.19D-F). Among other mature miRNA producing AGO-associated mRNA stem-loops were regions of the *Cyr61* 5' UTR, *Asf1b* 5' UTR, and *Klf9* 3' UTR (Figure 4.20). Collectively, these data suggest that pre-miRNA-seq uncovers AGO-associated mRNA stem-loops, some of which are likely to represent novel miRNAs, but most of which are poorly processed into AGO-bound smRNAs.

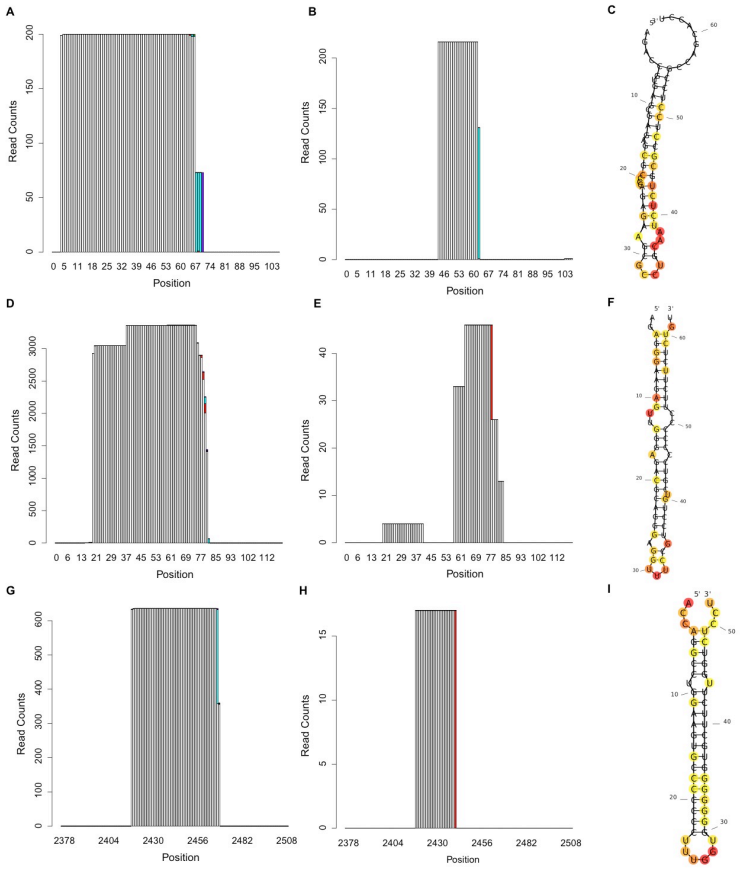


Figure 4.20 Novel miRNAs from AGO-associated stem-loops in mouse. A-I) Coverage plots of pre-miRNA-seq (A, D and G) and miRNA-seq (B, E and H) reads and RNAfold predicted structures overlaid with icSHAPE reactivity (C, F and I) for mouse AGO-associated stem-loops in the 5' UTR of *Cyr61* (A-C), 5' UTR of *Asf1b* (D-F) and 3' UTR of *Klf9* (G-I) White bars indicate templated nucleotides, colored bars indicate non-templated additions.

4.2.10 Iron response elements are processed into AGO-associated stem-loops

The top AGO-associated mRNA stem-loop candidate in humans is localized in the 5' UTR of Ferritin heavy chain (*FTH1*) (Figure 4.21A). We identified over 5,000 pre-miRNA-seq reads in this region, which had a strong predicted hairpin structure, clearly defined ends, and significant mono-tailing on the 3' end. Intriguingly, this region of the *FTH1* transcripts corresponds precisely to the iron responsive element (IRE), which is a well-studied structural element that regulates translation in an iron-dependent fashion [204] . We scanned our list of AGO-associated stem-loops for other IRE containing genes, and found Ferritin light chain (*FTL1*) was also producing AGO-associated stem-loops from its IRE region (Figure 4.21D)). Furthermore, we identified AGO-associated stem-loops supported by icSHAPE data from the mouse homologs of both of these genes, *Fth1* and *Ftl* (Figure 4.22). Therefore, processing of IREs from ferritin genes into AGO-associated stem-loops is conserved in mammals.

To examine whether IRE-processed hairpins produce functional small RNAs, we examined miRNA-seq and smRNA-seq data from these regions. We did not find any AGO-interacting small RNAs from human *FTH* or *FTL1*. We did however; observe cellular smRNAs from these regions in our smRNA-seq data (Figure 4.21 B,E**)**. However, they were heterogeneous in size and only loosely reflective of DICER processing (Figure 4.21C,F). Therefore, these smRNAs appear to be the consequence of subsequent degradation of the processed stem-loops that are not loaded into AGO to make a functional RISC complex. In mouse, miRNA-seq data we were only able to find a small number of clones originating from the *Fth* pre-miRNA (Figure 4.22D), further corroborating our results from human cells. Intriguingly, siRNA-mediated knockdown of *DROSHA* or *DGCR8* in HEK23T cells had no affect on *FTH1* or *FTL* mRNA expression levels (Figure 4.23). This suggests that processing of IREs into stem-loops is DROSHA-independent and may work through a different endonuclease. In total, these results demonstrate that mammalian IREs are processed into AGO-bound stem-loops through a microprocessor-independent mechanism, and that these stem-loops are not substrates for DICER. Whether cleaved IRE stem-loops are functional remains to be determined.

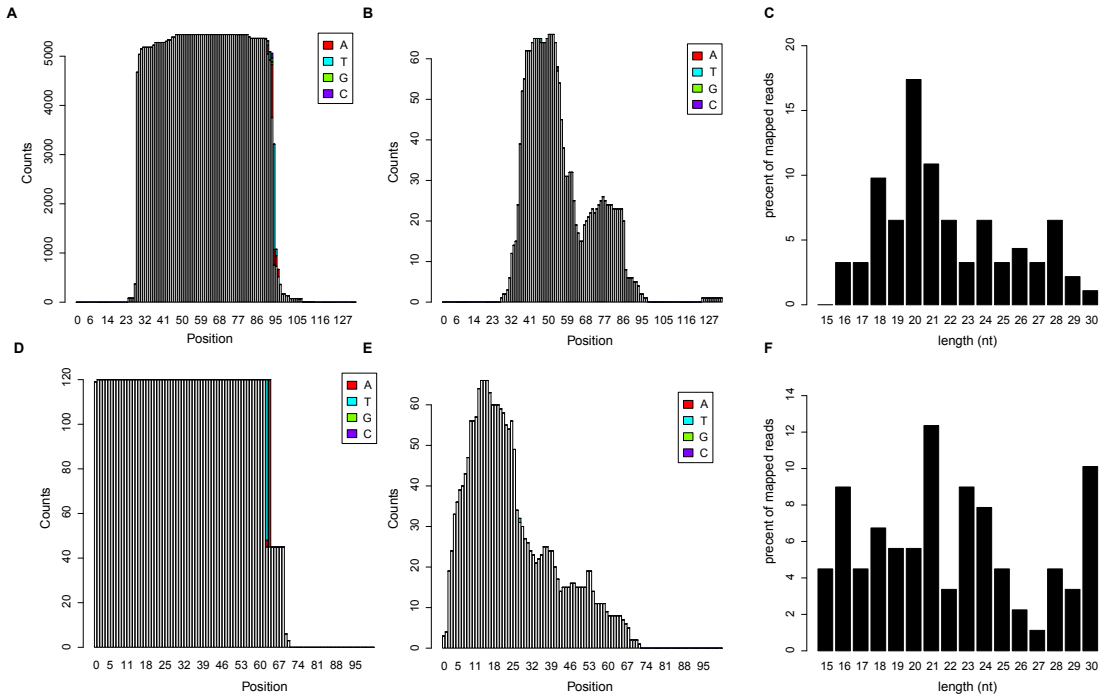


Figure 4.21 The IREs of human FTH1 and FTL are processed into AGO-associated stem-loops. A-B) Coverage plot of pre-miRNA-seq (A) and smRNA-seq (B) reads mapping to a human AGO-associated stem loop in the 5' UTR of the *FTH1* gene. White bars indicate templated nucleotides, colored bars indicate non-templated additions. C) Size distribution of smRNA-seq reads mapping to the human AGO-associated stem-loop in the 5' UTR of *FTH1*. D-E) Coverage plot of pre-miRNA-seq (E) and smRNA-seq (F) reads mapping to a human AGO-associated stem loop in the 5' UTR of the *FTL* gene. White bars indicate templated nucleotides, colored bars indicate non-templated additions. F) Size distribution of smRNA-seq reads mapping to the human AGO-associated stem-loop in the 5' UTR of *FTL*.

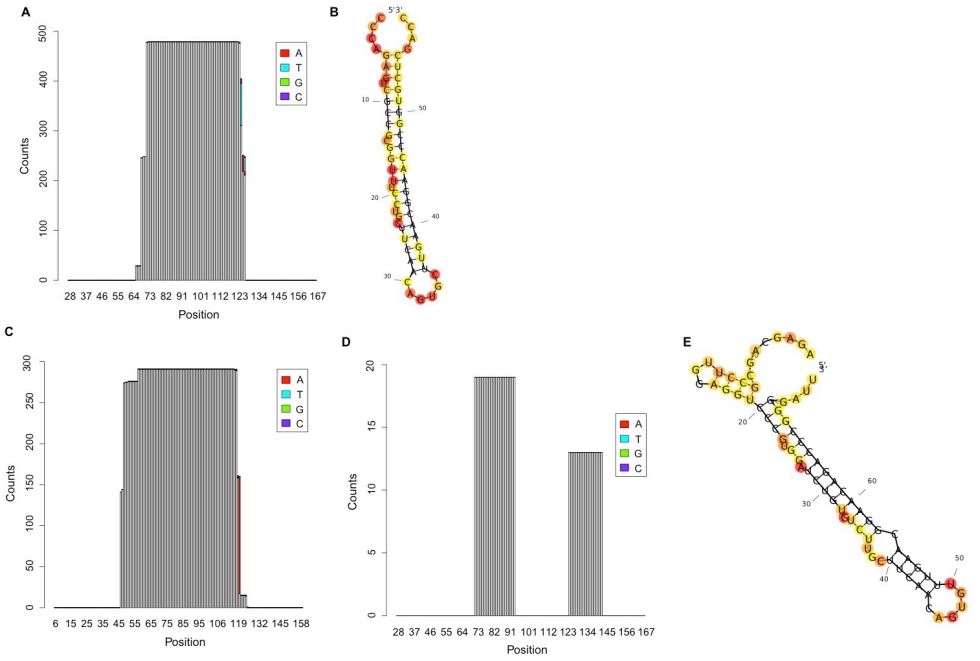


Figure 4.22 The IREs of mouse Fth1 and Ftl1 are processed into AGO-associated stem-loops. A-F) Coverage plot of pre-miRNA-seq (A and D) and miRNA-seq (B and E) reads and RNAfold predicted structures overlaid with icSHAPE reactivity (C and F) for mouse AGO-associated stem loop in the 5' UTR of the mouse *Fth1* (A-C) and *Ftl1* (D-F) mRNAs. White bars indicate templated nucleotides, colored bars indicate non-templated additions.

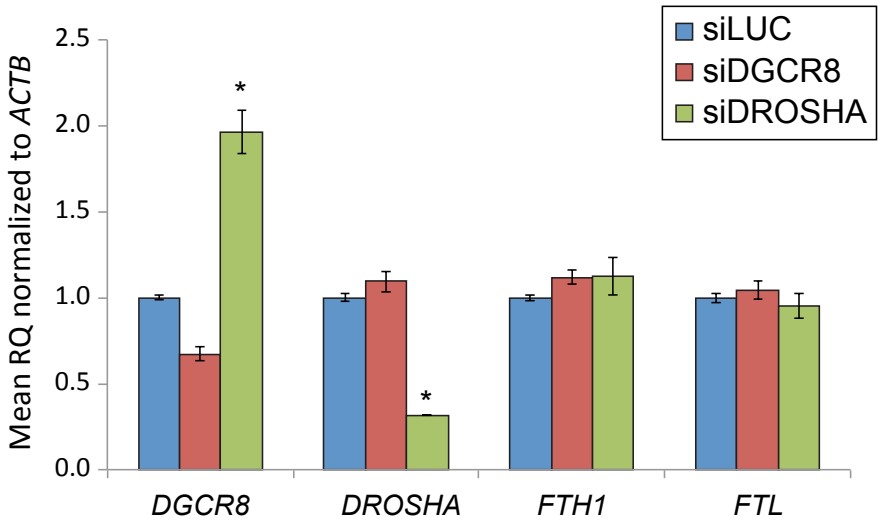


Figure 4.23 IRE host genes are unaffected by knockdown of microprocessor components. RT-qPCR analysis of *DGCR8*, *DROSHA*, *FTH1* and *FTL* following knockdown of indicated mRNAs. * = p-value <0.05,** = p-value <0.01, *** = p-value<0.001; Students' t-test.

4.3 DISCUSSION

Here, we describe the application and further development of a methodology to enrich for and sequence AGO-associated pre-miRNAs in both human and mouse cells. This biochemical approach combined with custom bioinformatics pipelines, successfully enriches for and maps pre-miRNAs in mammalian genomes (Figure 4.1-4.3). Using this approach, we detected 367 pre-miRNAs in human and 267 in mouse cell lines, with ~ 28 and ~ 17 million raw sequencing reads in each experiment, respectively (Table 4.1). This gave us specific insights into the exact sequence and abundance of pre-miRNAs and miRNAs expressed in cells of two different mammalian organisms. We generated profiles to visualize coverage, trimming and non-templated tailing at each annotated miRNA expressed in either cell type, which is available for download at http://gregorylab.bio.upenn.edu/AGO_IP_Seq/ (Example in Figure 4.4).

4.3.1 Insights into pre-miRNA processing

Using these unique datasets, we uncovered widespread trimming and non-templated tailing in both pre- and mature miRNAs from human and mouse cells (Figures 4.5-4.7). We also identified known and putative ac-pre-miRNAs (Figure 4.8A-B and Table 4.2). The large number of ac-pre-miRNAs identified suggests that DICER-independent pre-miRNA processing may be a more commonly used mechanism than previously appreciated [77, 78, 183]. Furthermore, we identified putative ac-pre-miRNAs that cleave in the 5p arm of the pre-miRNA (Figure 4.8C-D), and thus are processed in the opposite direction of the currently known members of this pre-miRNA class. This potentially novel pre-miRNA processing mechanism would require an alternative maturation process, with processive exonucleolytic nucleotide removal occurring step-wise from the 5' end.

Given the unique nature of our datasets, we were able to make the first comprehensive analysis of the relationship between pre-miRNA and mature miRNA abundance in an unbiased fashion (Figure 4.9). Remarkably, we found very consistent relationships between human and

mouse for the processing efficiencies of miRNAs. Using this unique approach, we determined a microRNA precursor processing index (MPPI), allowing us to determine productive and unproductive miRNA maturation (Figure 4.10**)**. We uncovered some pre-miRNAs that make surprisingly few mature species, despite abundance precursors. We also found that two miRNAs in Dgcr8 mRNA are processed with highly divergent efficiencies, suggesting distinct functionalities (Figure 4.11).

4.3.2 Identification of Cleaved AGO-associated stem-loops in mRNAs

From a further examination of pre-miRNA-seq reads that did not map to miRBase, we identified AGO-associated stem-loops that map to the exons of mRNAs (Figure 4.12). We found that these were enriched in the 5' UTR of these transcripts, which suggests they may play a similar role to miR-3618 in the 5' UTR of human Dgcr8 (Figure 4.13A-B). Furthermore, we found that these stem-loops had a broader size distribution than known miRNAs and were oligo-uridylated, suggesting they are processed by TUTases (Figures 4.13C-D and 4.14). RNA structure prediction algorithms and *in vivo* structure probing data from mouse ES cells provide strong evidence that these regions of mRNAs form stem-loops (Figure 4.15). Furthermore, AGO-bound stem-loop containing mRNAs were significantly upregulated following siRNA-mediated knockdown of components of the microprocessor complex (Figure 4.16). Collectively, our results strongly suggest the presence of AGO-associated stem-loops from mRNAs and that at least a subset of these interact with components of the miRNA biogenesis pathway.

Interestingly, we found very few mature AGO-bound sequences coming from these regions and overall low MPPI scores (Figures 4.17-4.18). In human, the vast majority of AGO-bound smRNAs from these regions can be explained by recently identified novel miRNAs [187, 202]. In mouse, we uncovered a region in the *Rpl9* CDS and a few other mRNA transcripts that account for most of the AGO-bound smRNAs from these regions (Figure 4.19). These findings indicate that our methodology uncovers novel miRNAs (Figure 4.20), but raises the question of

functionality of most of these stem-loops, which do not produce AGO-bound miRNAs, remains a mystery.

4.3.1 IREs are cleaved and AGO-bound

We also found that the IRE elements of human and mouse ferritin mRNAs are processed into pre-miRNA-like molecules, but not into mature AGO-bound miRNA species (Figure 4.21-4.22). Moreover, knockdown of *DROSHA* or *DGCR8* had no effect on the expression of IRE host genes (Figure 4.23). Then, what is the function of these pre-miRNA-like molecules? IRE hairpins may represent stable remnants of normal degradation of IRE containing mRNAs. However, given the conservation of AGO-associated stem-loops from IREs in both human and mouse cell, this seems unlikely. Alternatively, they may be processed from ferritin host genes by endonucleases other than the DGCR8/DROSHA microprocessor. This would be consistent with the inability of these pre-miRNA-like molecules to serve as substrates for DICER processing. However, these stem-loops could serve alternative roles, such as acting as a RNA-binding protein sink for IRE-binding proteins. Undoubtedly, the biogenesis and biological relevance of these processed stem-loops will be the subject of further investigation.

4.4 METHODS

Cell culture

HEK293T cells were grown to 70-80% confluence in 15 cm tissue cultures plates with DMEM media supplemented with 10% FBS and 1X pen/strep at 37°C and 5% CO_2.

AGO-IP-sequencing

Pre-miRNA-seq and miRNA-seq was performed as previously described [183]. Briefly, HEK293T cells were lysed in RSB 200 (20 mM Tris-HCl pH 7.4, 200 mM NaCl, 2.5 mM $MgCl_2$,

0.5% NP-40) supplemented with 0.2 U/μl RNaseIN (Promega) and 1 tab/10 ml of protease inhibitor cocktail (Roche) with 1-3 10 second bursts of sonication. Cleared lysates were incubated with Agarose protein G beads (Life Technologies) conjugated to 12 μg of AGO 2A8 antibody or 3 μl of non-immune serum (NIMS) per 1 ml of lysate. Conjugated beads were incubated with lysate for 1.5 h at 4°C on rotator and washed 4X with RSB 200. 500 μl of Trizol was added to washed beads and vortexed for 30 seconds. 150 μl of chloroform was added and the reaction and vortexed for 30 seconds followed by 20 min of centrifugation at 16,000 g. The supernatant was transferred to a new tube with 300 μl of isopropanol and 15 μg of glycogen. Pellets were recovered and RNA was dephosphorylated with phosphatase and labeled with T4 polynucleotide kinase and P^{32}-g-ATP.

5' end radiolabeled RNA from AGO-IPs were resolved on a 15% denaturing PAGE gel with 7 M urea. Gel slices from 20-25 nt (miRNA-seq) and 50-80 nt (pre-miRNA-seq) were recovered and ligated to miRCat 3' Linker (IDT) with T4 RNA Ligase 2 Truncated (NEB). Ligation products were resolved on a 15% PAGE gel with 7 M urea, size selected and purified. Reverse transcription was performed and product was purified from a 10% PAGE gel slice. cDNA was circularized using CircLigase I and PCR amplified. PCR amplicons were gel purified on a 3% Metaphor Gel, size selected and a second round of PCR was performed. Product was again size selected, purified, and submitted for sequencing.

Mapping Pipeline

The first 20 nt of the 3' adapter sequence *CTGTAGGCACCATCAATAGA* was used to trim adapter sequence from the raw reads using cutadapt (v1.4.2). Identical reads were collapsed but clone information was retained to reduce computational time. Trimmed reads were sequentially mapped to miRBase (v20) and RefSeq annotated spliced transcript models (hg19 or mm10, downloaded on 06082015), with a two-stage alignment strategy with Bowtie2 and

EMBOSS-WATER. First, Bowtie2 was used to map the 5' regions of reads to either miRbase (v20) primary miRNA sequences with a 50 nt extension on the 3' end or RefSeq annotated spliced transcript models. For pre-miRNA-seq, the first 35 nt were used in the initial alignment step, whereas 18 nt were used for miRNA- and smRNA-seq. Following Bowtie2 alignment, reads were extended by local alignment with EMBOSS-WATER (with parameters: -gapopen 10.0 -gapextend 0.5). Multimapped reads were partially resolved by selecting the longest, highest scoring alignments. Mismatches detected at the 3' ends of reads were considered as non-templated additions and analyzed separately. We filtered unmapped reads from pre-miRNA-seq for known rRNA, snoRNA, snRNA, tRNA, mitochondrial transcripts, and repeat-masked sequences before mapping to RefSeq.

Analysis of miRNA trimming and non-templated tailing

We tabulated non-templated additions revealed by our mapping pipeline for reads, which mapped to high-confidence miRBase annotations. Additionally, we calculated templated-extension and trimming by comparing mapped ends of pre-miRNA-seq, miRNA-seq, and smRNA-seq reads against annotated 5p and 3p miRNA ends.

Identification of AGO-2 cleaved pre-miRNAs

For each miRBase miRNA, we calculated the percentage of trimmed pre-miRNA-seq reads that ended between 8-15 nts upstream of the 3' end of the 3p miRNA or downstream of the 5' end of the 5p miRNA. We identified pre-miRNAs with >1% of such cleavage events in the 3p arm as putative ac-pre-miRNAs. For cleavage events in the 5p arm, we took a more conservative approach, requiring 5% of reads to terminate in this region and > 50% of mature miRNA-seq reads in the 3p arm compared to 5p arm.

Determination of MPPI

We calculated the miRNA precursor processing index (MPPI) for miRBase miRNAs or AGO-associated stem-loops as the generalized log ratio (glog) [129, 205, 206] of miRNA-seq (smRNA-seq) to pre-miRNA-seq RPM coverage (nmi, npr) as follows:

$$MPPI_i = glog(nmi_i) - glog(npr_i) = log_2\left(nmi_i + \sqrt{1 + nmi_i^2}\right) - log_2\left(npr_i + \sqrt{1 + npr_i^2}\right)$$

Thus MPPI can be calculated for loci that are not represented in either the pre-miR-seq or miR-seq library. Loci with no coverage in either sequencing library were omitted from this analysis.

Identification of AGO-associated stem-loops in mRNAs

Pre-miRNA-seq reads that failed to map to miRBase and were not removed by mapping to known rRNA, snoRNA, snRNA, tRNA, mitochondrial transcripts and repeat-masked sequences were we aligned to RefSeq transcripts with Bowtie2. After filtering for best matches for reads with less than 4 mismatches, we used Pycioclip to call significant peaks with a modified False Discovery Rate < 0.01 [151]. To remove abundant genes with high numbers of mappings but no local peaks, we filtered out peaks that were greater than 200 nt in length. We then chose the most abundant clone and predicted its secondary structure with RNAfold using standard parameters [201]. We again filtered clusters, requiring that they had greater than 15b p in the longest hairpin and a total MFE of less than 0.3 kcal/mol/nt. We analyzed non-templated tailing for reads mapping to these hairpins as described above.

icSHAPE analysis

In vivo icSHAPE data from mouse ES cells was downloaded from GSE60034. Only scores in the top or bottom 10^{th} percentile were used for analyisis. icSHAPE reactivity scores were overlaid on RNAfold diagrams using RNAplot [201].

siRNA-mediated knockdown of DGCR8 and DROSHA

ON-TARGET plus siRNAs against Human *Dgcr8 and Drosha* were obtained from Dharmacon (J-015713-05-0002, J-015713-06-0002, J-016996-05-0002, J-016996-06-0002). siRNAs against luciferase were a gift from the Mourelatos lab. To knockdown endogenous levels of *Dgcr8* or *Drosha*, we performed two sequential siRNA transfections 48 hours apart. To transfect these cells, we combined 45 pmol of siRNAs (22.5 pmol siRNA-1 and 22.5 pmol siRNA-2), 125 mL Opti-MEM (Life Technologies), and 6 mL Lipofectamine 2000 (Life Technologies) per reaction. Reactions were incubated at room temperature for 20 minutes. During this incubation, we seeded $6.0x10^5$ HEK293T cells per replicate in two wells of a 6-well plate ($3.0x10^5$ cells/well) in 2 mL media. We then added the siRNA mixture to each well dropwise, and allowed cells to incubate at 37ºC in 5% CO_2 for 48 hours. After 48 hours we repeated the transfection, harvesting the two wells per replicate (~$1.2x10^5$ cells) and dividing them into four wells in a 6-well plate. We treated each well with 45 pmol of siRNAs, and allowed them to incubate at 37ºC in 5% CO_2 for another 48 hours. Cells were then pooled and washed with PBS prior to storage at -80ºC.

mRNA-seq

mRNA-seq was performed as previously described [172] . Briefly, total RNA was purified from the MEL cell cultures (miRNeasy; Qiagen, Valencia, CA). Poly(A)+ RNA was isolated using oligo dT beads (Life Technologies, Frederick, MD). RNA was fragmented for 7 minutes using

Fragmentation Reagent (Life Technologies, Waltham, MA). mRNA-seq libraries were then generated using the Illumina mRNA-seq kit (illumina, San Diego, CA). Reads were trimmed with Cutadapt, mapped with Tophat2, and gene expression was quantified using HTseq [173-175]. DEseq2 was used perform differential expression analysis [207].

Data access

All sequencing data generated in this study has been deposited in GEO under the accession number GSE71710 (available at http://www.ncbi.nlm.nih.gov/geo/). RNAfold diagrams and coverage plots are available for download at http://gregorylab.bio.upenn.edu/AGO_IP_Seq/. MEF pre-miRNA-seq and miRNA-seq data were downloaded from European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under the accession number PRJEB6756. HEK293T smRNA-seq data was obtained from GSE66224.

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS

Abstract:

In this dissertation, I have covered a diverse array of topics connected by the common theme of RNA-protein interactions. In Chapter 2, we developed a novel methodology for uncovering RBP-RNA interaction sites throughout a transcriptome of interest. In Chapter 3, we generated a transcriptome-wide map of RNA binding sites for a single RBP, PABPC1. In Chapter 4, we leveraged the interaction of pre-miRNAs with AGO proteins to sequence these transient intermediates and study them in two organisms. The variety of studies performed here underscores the range of biological pathways and regulatory roles for which RNA-protein interactions are key components. In this section, we will discuss the major advances provided by each of these studies and propose future experiments to address questions that have arisen from this work.

5.1 A NOVEL APPROACH TO IDENTIFY RNA-PROTEIN INERACTION SITES

In Chapter 2, we introduced a novel methodology, PIP-seq, to uncover RNA-protein interaction sites throughout a transcriptome of interest. This technique represents a significant advance over previous technologies, which only probe the RNA binding sites of a single protein at a time. PIP-seq is distinct from related techniques that were introduced concurrently because it does not rely on the use of synthetic nucleotides, which cannot be used in tissues or whole organisms [22, 106]. We applied PIP-seq to uncover protein-bounds sites in HeLa cells and provided multiple lines of evidence to support the accuracy and applicability of this novel methodology [208]. Importantly, we showed that PIP-seq is both reproducible and identifies previously identified sites of RBP interactions in mRNAs. Therefore, we now have a validated new tool in our arsenal to explore RBP binding sites on mRNAs.

5.1.1 RBP Occupancy Profiles on mRNA and lncRNA

We used PIP-seq to investigate aspects of RNA-protein interactions that were not previously addressed due to limitations of earlier methods. First, we asked what the global landscape of RBP binding across mRNAs looked like. We found that the 3' UTR and CDS were more bound than the 5' UTR (Figures 2.12 and 2.15). Although we can't distinguish PPSs within the CDS from ribosome occupancy, the increased binding in the 3' UTR as compared to the 5' UTR confirms years of evidence that most RBPs interact with regions in the 3' UTR RBP 'sanctuary'. We also directly compared protein binding at long non-coding RNAs (lncRNAs) and expression matched mRNA 3' UTRs (Figure 2.13). We found similar amounts of PPSs between these two RNA types when using formaldehyde as a cross-linking reagent. However, lncRNAs were depleted when using UV cross-linking, which only crosslinks direct RNA-protein contacts. This suggests that lncRNAs serve as platforms for RBP binding, but that these interactions may be relatively weak, consistent with the low sequence conservation observed across lncRNAs. Together, these results confirm previous narrowly focused studies and offer new insights into the mechanisms of lncRNA-mediated gene regulation, which have been thought to act as RNA scaffolds for protein and DNA interaction [209]. Future experiments will be focused on further characterizing lncRNA-protein interactions and understanding what distinguishes them from mRNA-protein interactions.

5.1.2 Insights into RNA Regulons

Using our PIP-seq data, we explored the post-transcriptional operon, or regulon hypothesis of mRNA regulation in eukaryotes. This hypothesis states that mRNAs involved in the same functional pathways are regulated by similar sets of RBPs, in a manner akin to prokaryotic operons [210, 211]. Our analyses demonstrated that certain putative RBP-interacting motifs tended to co-occur on the same transcripts, and that some of these groups of transcripts fell into pathways of immune regulation, RNA production, and cell death (Figure 2.16). It would be

interesting to take this work further by identifying putative-motif interacting RBPs by RNA affinity chromatography [108]. We propose performing this analysis in other cell lines or primary tissues to identify cell-type specific pathways that are controlled by regulons. Knockdown or knockout studies coupled to RNA-seq or phenotyping could be used to confirm that these RBPs are key regulators of predicted processes.

5.1.3 Insights into Human Disease

We also used PIP-seq data to learn something about human disease. We found that disease-linked single nucleotide polymorphisms (SNPs) were enriched within PPSs (Figure 2.17). Furthermore, we found that synonymous SNPs, which change DNA/RNA sequence but not primary protein sequence, were enriched compared to nonsynonymos SNPs within PPSs. We validated the ability of two such SNPs to alter protein binding to specific mRNAs. These SNPs could impact on RNAs in numerous ways, including any of the post-transcriptional mechanisms described in the introduction. Collectively, these data suggest that disruption of RNA-protein interactions may be a more common mechanism for human disease than previously thought. We propose that PIP-seq could be performed in disease-relevant tissues to more accurately identify specific SNPs that disrupt RNA-protein interaction sites. Those interaction sites that overlap with disease-linked SNPs could be flagged as potential RNA-protein interaction disruptors, which would enable researchers to more efficiently investigate the mechanisms of some diseases.

5.1.4 PIP-seq in Plants

Since the development of PIP-seq (Chapter 2), our laboratory has applied it to study RBP-interaction sites in the HEK293T transcriptome and nuclear RNAs in the flowering plant, *Arabidopsis thaliana* [108, 212]. Ongoing studies in our lab are also characterizing RNA-protein interaction sites in the bryophyte *Physcomitrella patens* (moss) and in *Zea mays* (corn). PIP-seq is of particular importance to the study of plants, for which RBPs remain relatively

uncharacterized. Performing PIP-seq on nuclear fractions was an advance over our original approach, because we could not previously distinguish between ribosome occupancy and RBP binding within the CDS.

5.1.5 Dynamic RNA-Protein Interactions

One of the main advantages of PIP-seq is that it can monitor the expression of a large number of RBP binding sites in a single experiment. How do RBP-RNA interactions change over time in dynamic biological processes and how are these changes reflected in post-transcriptional regulation? The long-term goal of our lab is to answer these questions and PIP-seq offers us a means to address them. We are currently exploring this question in a mouse model of blood cell development. Erythropoiesis is an ideal system for this type of analysis because blood cells function without active transcription and rely exclusively on post-transcriptional controls [213]. We have performed PIP-seq on mouse erythroleukemia (MEL) cells at three time points during DMSO induction of differentiation; Day 0 (uninduced), Day 2 and Day 4. We have developed a computational pipeline to identify RNA-protein interaction sites that show significant changes in occupancy during the differentiation process, utilizing the DESeq package [214] applied pairwise (e.g 0 vs. 2, 2 vs. 4, and 0 vs. 4) to pseudo-read counts. These pseudo-read counts will be calculated in a manner to preserve the relative read ratios of the Footprinting (Fp) sample to the digestion control (Dc) samples within each time point. Specifically, we will compute pseudocounts for each j-th PPS under conditions a and b: $C_j^a = [r_i^{a(Fp)}/r_i^{a(Dc)}] * R_i$ and $C_j^b = [r_i^{b(Fp)}/r_i^{b(Ds)}] * R_i$ where $r_i^{a(Fp)}$ are the read coverage from Footprinting (Fp) samples and $r_i^{a(Dc)}$ are the read coverage from the Digestion control (Dc) samples of the j-th PPS under condtion a. The ratio $r_i^{a(Fp)}/r_i^{a(Dc)}$ is thus a vector (across replicates) of the relationship between the Footprint and Digestion control sample read coverage under condition a. $r_i^{b(Fp)}/r_i^{b(Dc)}$ is a vector (across replicates) of the relationship between Fp and Dc read coverage under condition b, and R_j is the average read coverage across all Dc replicates in both conditions. We can then directly compare the pseudocount vectors C_j^a

and C_j^b to identify differentiation-impacted RNA-protein interaction sites. This analysis pipeline will enable us to identify differentially bound regions during erythroid differentiation.

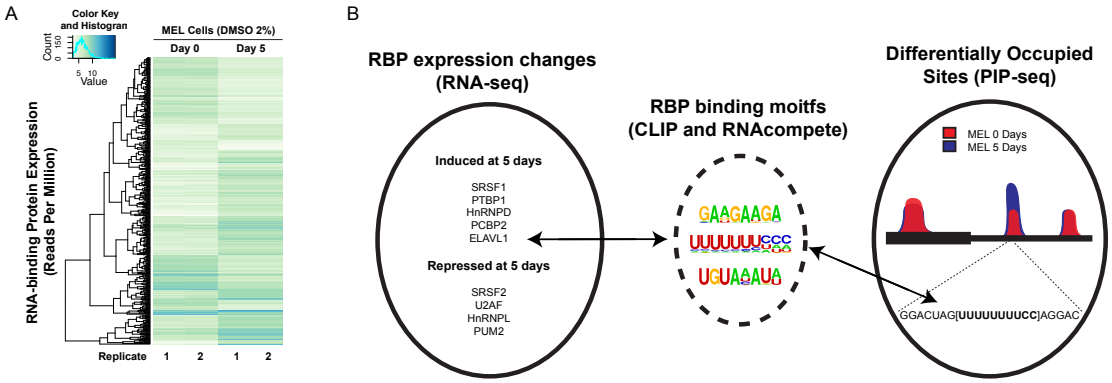


Figure 5.1 Intersection of RNP expression profiles, motif libraries, and PIP-seq data. A) Heatmap showing dramatic shifts in the RBP profiles during MEL cell induction (Day 0 → 4). Hierarchical clustering analysis was performed on RBP expression changes (y-axis) to reveal distinct groups of similarly regulated RBPs. B) Example: PIP-seq data set (right circle) identifies complex enrichment at a site within an RNA 3' UTR at Day 4. The core PPS at this site (center circle) matches the consensus motif for the RBP, ELAVL1. Expression of ELAVL1 is increased at Day 4 of differentiation as inferred from RNA-seq analysis shown in A). These theoretical informatic comparisons identify ELAV1 as a candidate binding protein at this site.

It will be interesting to identify shifts in protein occupancy on RNA using PIP-seq and the computational pipeline described above, in combination with other approaches to determine proteins responsible for these changes (Figure 5.1). For example, we propose using a combination of RNA-seq measurements with the RNAcompete compendium of RNA-binding motifs [94, 95] to predict which RBPs are responsible for interactions at each site. We can then identify RBPs, whose occupancy profile is most changed during differentiation and define their impacts on gene expression or other post-transcriptional processes through global measurements of alternative splicing, translation and/or RNA stability. We already know that the occupancy of PCBP2 within globin mRNA increases during differentiation and that this binding stabilizes globin mRNAs [215]. Thus, PCBP2 will serve as important positive control when performing such analyses. For novel candidates, CLIP-seq and/or RIP-qPCR could then be used

to validate these interactions and their dynamics in blood cell differentiation. Follow up studies using siRNA knockdown can also be used to evaluate the impact of loss of individual RBPs on blood cell development. Application of PIP-seq to erythropoiesis and other dynamic biological systems will likely uncover RBPs with important roles in these processes, which has remained difficult despite the number of technologies available.

5.2 EXPANDING ROLES FOR PABPC1 IN GENE REGULATION

In Chapter 3, we identified the genomically encoded RNA targets of mammalian PABPC1 throughout the mouse transcriptome. This was the first high-resolution global analysis of the binding sites of any mammalian PABP. Using this approach, we uncovered three distinct modes of binding to mRNAs outside of its known role in poly(A) tail binding. First we revealed that PABPC1 binds directly to the polyadenylation signal (PAS) of thousands of mouse mRNA transcripts. We also found that PABPC1 binds to the start and stop codons of mRNAs, in the absence of any underlying sequence motif. Finally, we showed that PABPC1 binds to the A-rich elements in the 5' UTR of numerous mRNAs and negatively regulates their translation. Here, we discuss the implications for these findings and delineate future experiments to address new questions that have arisen from these novel insights.

In our study, we found that CLIP tags were highly enriched in the 3' UTR and specifically localized towards the 3' terminus of mRNAs (Figure 3.3). Unfortunately, the length of the sequenced reads limits the resolution of CLIP-seq. Classically, peak identification is performed by identifying clusters of overlapping reads to define broad binding sites. However, it was recently noticed that CLIP tags are specifically enriched for deletion events relative to other types of sequencing technologies [146, 147]. It is thought that these deletions are a result of protein-fragments that remain on the RNA following proteinase K treatment. Therefore, these deletion events can be leveraged to identify RNA-protein interaction sites with singe-nucleotide resolution.

We applied this approach, termed cross-link induced mutation site (CIMS) analysis, to precisely identify PABPC1 interaction-sites within the murine transcriptome (Figure 3.4).

5.2.1 PABPC1 Binds Directly to the Polyadenylation Signal

Using CIMS analysis, we found that PABPC1 binding events were most enriched approximately 20-25 nucleotides (nt) upstream of the poly(A) addition site (Figure 3.7). This region coincides with the precise location of the mammalian polyadenylation signal sequence (PAS), which functions in the nucleus to recruit CPSF, a core component of the polyadenylation machinery [216]. CPSF also binds to the PAS in the cytoplasm to execute a much less common mechanism of cytoplasmic polyadenylation [217]. We also performed motif enrichment analysis of sequences proximal to PABPC1 CIMS sites and found that the PAS sequence (AAUAAA) was the most enriched sequence motif (Figure 3.6). Finally, we demonstrated that active polyadenylation sites were ~25 nt downstream of CIMS sites, again supporting the notion that PABPC1 binds directly to these regions (Figure 3.7). This is the first report of mammalian PABP interacting with the PAS, however, a related interaction was observed at the yeast efficiency element, which also plays a role in polyadenylation [59, 61].

Given that PABPC1 has been previously shown to bind to A- and AU-rich sequences, it is not surprising that PABPC1 can bind to the PAS sequence [47]. This interaction would effectively extend the region of the mRNA protected by PABP further upstream from the poly(A) tail. We did not determine the functional consequences of this binding, but we hypothesize that it would increase the stability and translation efficiency of bound mRNAs. In order to test for functionality of this interaction, careful experimental design must be taken, given the defined roles of PABPC1 in regulating translation and RNA stability, as well as the PAS in polyadenylation. To circumvent these problems, we propose directly transfecting pre-polyadenylated luciferase mRNAs with or without the PAS signal to bypass polyadenylation in the nucleus. Ideally, we would knockdown PABPC1 levels with siRNAs, however, the global affects of this loss may preclude identification of

the function of PABPC1-PAS binding. We would then assay mRNA stability and translation over time using the dual luciferase reporter system. We expect that additional PABPC1 binding to the PAS sequence would stabilize the mRNA and promote translation. Alternatively, binding of PABPC1 to the PAS in the cytoplasm may block binding by CPSF and subsequence cytoplasmic polyadenylation. Competition assays between CPSF and PABPC1 binding, or tethering of CPSF to mRNA 3' UTR in the absence of the PAS could be used to address this question.

5.2.2 PABPC1 at Translation Initiation and Termination Sites

We also observed that PABPC1 interacts with sites of translation initiation and termination in mRNAs (Figures 3.8-3.9). This was most dramatically observed in the replication-dependent histone mRNAs, which are unique among mRNAs in that they lack a poly(A) tail. We were unable to identity an enriched sequence motif in these regions, suggesting that these interactions are sequence-independent. An explanation for the presence of PABPC1 at these regions may lie in its known interactions with other factors. Specifically, PABPC1 binds to EIF4G, which in turn binds to EIF4E, which interacts with the ribosome to activate translation [218]. This interaction can circularize the mRNA, however, it has not been clearly demonstrated that PABPC1 remains associated with this complex during active translation [53]. Furthermore, replication-dependent histone mRNAs lack a poly(A) tail and thus circularization would not be predicted to involve PABPC1 at these mRNAs. Our observations suggest PABPC1-EIF4G-EIF4E-ribosome interactions are stable at both the start and stop codons. It's possible that release factors would be required to break this interaction to begin translation or promote ribosome release.

To experimentally address the mechanism of PABPC1's interaction at translation initiation and termination sites, we propose genetic ablation of the PABPC1-EIF4G interaction domain and CLIP-seq on PABPC1. If peaks at the CDS start and stop codon are lost, this would suggest that these interactions are mediated through EIF4G. This assay may be challenging due

to global impacts on translation from loss of the PABPC1-EIF4G interaction. It would be interesting to compare the effects on translation due to loss of this interaction for histone mRNAs relative to polyadenylated mRNAs. If the role of this interaction is mRNA circularization, then histone mRNAs should be unaffected.

5.2.3 PABPC1 Binds to and Regulates Specific mRNAs

It has been previously observed that PABPC1 protein binds to an A-rich tract in the 5' UTR of its own mRNA transcript and negatively regulates translation [41, 58]. Our CLIP-seq data strongly support this finding and also showed that PABPC1 interacts with another A-rich region downstream of the known regulatory element region in its mRNA (Figures 3.12-3.13). We used an *in vitro* luciferase assay to show that the known PABPC1 binding site in the 5' UTR of *Pabpc1* mRNA acts as a negative regulator of translation. Surprisingly, we found the other A-rich region, with a more enriched binding peak, had no effect on mRNA translation or expression. This binding peak also happens to coincide with a predicted upstream open reading frame (uORF) with extensive experimental support from ribosome profiling with harringtonin treatment (Figure 3.14) [159]. We hypothesize that interaction in this region may regulate expression of the uORF but not the downstream main ORF. We also used the CRISPR/Cas9 system to genetically ablate the PABPC1 binding site in the 5' UTR of *Pabpc1* mRNA, in cells (Figure 3.15). This analysis showed that this A-rich tract is a repressor of translation but does not affect mRNA levels *in vivo.* Thus, we have provided extensive supporting evidence for the role of PABPC1 in auto-regulation through the 5' UTR of its own mRNA. However, our study further complicates this known interaction due to the presence of another more enriched PABPC1 binding site with an undefined function.

In addition to its interaction with the 5' UTR of *Pabpc1*, we also found that PABPC1 binds the 5' UTR of numerous other cellular mRNAs. This is the first such report of additional 5' UTR targets for a mammalian PABP. We performed luciferase reporter assays for 5' UTR target genes (*Ccnd2, Safb, and Amd1*), and found that the presence of PABPC1 binding sites has repressive

effects on translation (Figures 3.16-3.17). This is consistent with our studies of PABPC1 binding sites in the *Pabpc1* 5' UTR. However, we have now expanded this regulatory affect to multiple mRNA targets. There are hundreds of other mRNAs with PABPC1 binding sites in their 5' UTRs, and thus it is likely that PABPC1 also negatively regulates their translation.

One important caveat to these studies is that we showed that the PABPC1 binding sites and not PABPC1 itself is responsible for the translation inhibitory effects. Ideally, we would knockdown PABPC1 expression and assay for effects on translation of target genes, *in vivo*. However, we were not able to achieve PABPC1 knockdown without significant levels of cell death, and there are no reported PABPC1 KO mice to study. This is likely due to the global role of PABPC1 in mRNA stability and translation. Therefore, titrating knockdown conditions such that PABPC1 levels are reduced just ~10% could alleviate some of these problems. If we could achieve this, we would also perform a global assay for translation (ribosome profiling), so that we could assess ribosome occupancy on thousands of mRNAs simultaneously. We would expect that overall translation rates for PABPC1 5' UTR target genes would be increased. Despite lower overall translation in the system.

A recent study from the Liebhaber lab suggested that depletion of a minor PABP isoform, PABPC4, impacted on the maturation of erythoid cells *in vitro* [219]. We also performed CLIP-seq on PABPC4 and found that it shared the majority of its targets with PABPC1. PABPC4 is expressed at about 10% the levels of PABPC1 and therefore, it is reasonable to hypothesize that depletion of PABPC4 may be equivalent to a small reduction of PABPC1. Thus, defects in erythroid differentiation may be due to loss of PABP binding activity. PABPC4 knockdown does not result in cell death, thus, future studies will examine the impact of PABPC4 depletion on the translation of genes with PABPC1/C4 5' UTR binding sites.

5.3 NEW INSIGHTS INTO MICRORNA STEM-LOOP PROCESSING

In Chapter 3, we used a recently developed technique to isolate and sequence microRNA precursors (pre-miRNAs) in human HEK293T cells. We then developed a computational pipeline

to handle mapping of these highly modified RNA molecules which we applied to our data as well as data from mouse embryonic fibroblasts (MEFs), recently published by our lab [183]. Using these datasets we uncovered novel insights into the processing and post-transcriptional modification of pre-miRNAs. We also uncovered numerous AGO-associated stem-loops embedded within mRNAs, which are poorly processed into mature miRNAs. Here, we discuss the most important findings from our study and outline future experiments to better understand these results.

Our enrichment and sequencing strategy for pre-miRNAs represents a significant advance over earlier methods, which utilized size selection or primer-based amplification to sequence pre-miRNAs. Previous studies were unable to enrich pre-miRNAs to greater than 1% of sequenced RNA [188, 220]. Alternatively, primer-based approaches suffer from inherent bias and cannot be used for discovery [184]. In our study, pre-miRNAs represented ~10% of human and ~40% of mouse pre-miRNA-seq reads, providing detailed information on >600 mammalian pre-miRNAs (Figure 4.2). These data provide a wealth of novel information about the sequence and abundance of pre-miRNAs, which improve our understanding of these understudied intermediates. Application of this technique to other cell types, tissues and organisms could greatly enhance our annotations of miRNAs.

5.3.1 Diversity of pre-miRNA 3' ends

One of the most surprising results from our study was that the majority of pre-miRNA-seq reads did not map to the annotated 3' end of the pre-miRNAs. In MEFs only 20% of reads ended at the annotated terminal 3' position of mouse pre-miRNAs, while in HEK293T cells just 40% of reads ended at the expected 3' position (Figure 4.5). In contrast, >90% of pre-miRNA-seq 5' ends corresponded to the annotated 5' end of the pre-miRNA. We also found that 3' end trimming was more common than templated extension. These results demonstrate that pre-miRNAs in cells rarely end at the predicted 3' end position. There are several explanations for this finding. For

instance, Drosha cleavage could be more heterogeneous than previously thought, or other nucleases could “nibble” at pre-miRNA ends similar to what has been observed for mature miRNAs and ac-pre-miRNAs [181, 221]. Alternatively, pre-miRNAs with appropriate 3’ ends may be rapidly processed into mature miRNAs, and therefore do not accumulate in AGO proteins. Regardless of the explanation, these results show that we still have much to learn about pre-miRNA processing *in vivo.*

5.3.2 Identification of Novel ac-pre-miRNAs

We also leveraged our dataset to identify novel ac-pre-miRNAs, which are directly cleaved by AGO2 and subsequently turned into mature miRNAs by PARN [78-80, 181]. We identified all known ac-pre-miRNAs, with the exception of pre-miR-451, which was not expressed in either of the cell lines used in our study (Table 4.2). Furthermore, we identified numerous candidate ac-pre-miRNAs, which show similar 3’ cliff patterns (Figure 4.8). Follow up studies could be performed to demonstrate that these pre-miRNAs are indeed ac-pre-miRNAs. To do this, cells would be co-transfected with pre-miRNAs of interest, and AGO2 overexpression. Northern blots would be used to confirm that increasing levels of AGO2 results in increased levels of mature miRNAs. These results show that ac-pre-miRNAs are more widespread than previously thought.

To date, all known ac-pre-miRNAs are cleaved in the 3p miRNA to produce mature 5p miRNAs. We performed a parallel search for ac-pre-miRNAs that cleave in the 5p miRNA to produce 3p miRNAs. We found several candidates that show striking 5p cliffs and produce predominantly 3p mature miRNAs (Figure 4.8). To validate that these are 5p-ac-pre-miRNAs, we would perform the same validation assay described for novel 3p-ac-pre-miRNAs. If indeed these candidates are AGO2 cleaved, their maturation would require a distinct mechanism from what has been proposed for known 3p-ac-pre-miRNAs. A 5’ to 3’ exoribonuclease, for example XRN1 or XRN2, would be required for maturation from the 5’ end. I hypothesize that XRN1 is the likely

nuclease because XRN1 is cytoplasmic and AGO2 and XRN1 have already been shown to interact within cytoplasmic processing bodies [18, 222]. To test this hypothesis, *XRN1* could be knocked down with siRNAs and the levels of mature miRNA assayed for the putative 5p-ac-pre-miRNAs. From these experiments, I expect that mature levels would decrease in the absence of XRN1. In total, our studies have revealed a more complex landscape of AGO2-mediated miRNA maturation.

5.3.3 Insights into pre-miRNA Processing Efficiency

Given the comprehensive nature of our data, we were able to ask an additional basic, but unanswered question in miRNA biology; what is the relationship between pre-miRNA and mature miRNA abundance? One important caveat to this analysis is that numerous pre-miRNAs can give rise to the same mature miRNA species. Therefore, we took a conservative approach by first grouping miRNAs by family and then assessing the ratio of pre-miRNA and mature miRNA. We found a modest positive correlation (R = 0.53) that was surprisingly consistent between mouse and human annotated miRNAs (Figure 4.9). This suggested that while pre-miRNA and mature miRNA expression are related, that a significant amount of variation in this relationship exists between different miRNAs. To further explore this variation, we created an index based on the ratio of mature miRNAs to pre-miRNAs (MPPI). We observed that the majority of high confidence miRBase miRNAs scored greater than 0 on this scale. However, we identified some interesting outliers, defining efficiently and inefficiently processed miRNAs (Figure 4.10). Further investigation as to what differentiates these outliers from other miRNAs will likely lead to a better understanding of miRNA processing.

Among the poorly processed miRNAs was a miRNA encoded in the 5' UTR of the *DGCR8* mRNA. In fact, an adjacent pre-miRNA encoded in the CDS of *DGCR8* is efficiently processed into mature miRNAs (Figure 4.11). Collectively, these miRNAs have been demonstrated to negatively regulate the expression of *DGCR8* mRNA abundance through microprocessor-mediated cleavage [35, 199, 200]. However, it has not been investigated which

miRNA imparts this auto-regulatory function. To test this, I would generate luciferase reporter mRNAs with either miRNA inserted in the 5' UTR and assay for luciferase protein and mRNA levels. I expect that the poorly processed 5' UTR miRNA will impart the major regulatory potential, given that it appears to have no other function. However, the *in vivo* relevance remains convoluted, given that cleavage of either miRNA would lead to degradation, and likely cleavage of the other pre-miRNA.

5.3.4 Identification of pre-miRNA-like Elements in mRNAs

We further explored our dataset to identify other mRNAs with pre-miRNA-like sequences embedded in their exons and identified ~400 AGO-associated stem-loops in both mouse and human mRNAs. We provided compelling support with multiple lines of evidence for the validity of these elements (Figures 4.13-4.16). Furthermore, 12 of AGO-associated stem-loops have been recently identified as novel human miRNAs [187, 202]. However, beyond a few cases most AGO-associated stem-loops in human and mouse mRNAs did not produce AGO-bound smRNAs (Figure 4.18). Furthermore, these elements were highly enriched within the 5' UTR, suggesting they may be related to the *DGCR8* 5' UTR hairpin (Figure 4.13). Collectively, we have identified a set of stem-loop elements in mRNAs that are processed into AGO-associated pre-miRNA-like molecules, but fail to mature along the canonical pre-mIRNA pathway.

The most abundant AGO-associated stem-loop in our human dataset corresponded to the iron response element (IRE) of ferritin heavy chain (*FTH1*). Further investigation revealed that the IRE of ferritin light chain (*FTL*), as well as the mouse homologs of both of these elements are processed into AGO-associated stem-loops (Figure 4.21-4.22). We found that these elements do not produce AGO-bound smRNAs, but do produce a small number of cellular smRNAs. However, these unbound smRNAs were not arranged in a clean double stack and their size distribution did not correspond to Dicer cleavage. Finally, *FTH1* and *FTL* mRNA levels were not affected by *DROSHA* or *DGCR8* knockdown. These data suggest that IREs are cleaved in a microprocessor-

independent fashion into AGO-associated stem-loops, and thus are poor substrates for miRNA maturation.

What could be the function of these AGO-associated stem-loops, and specifically IRE-derived AGO-associated stem-loops? One possibility is that they are degradation products. The inherent stability of the stem-loop and perhaps AGO's affinity for it may lead to stabilization of this product after degradation. Alternatively, these may function as RBP sinks, interacting with and/or buffering the amount of specific RBPs available. In the case of IREs, this would correspond to the IRE-binding protein (IRE-BP), which mediate translation regulation through interaction with IRE elements [223]. The function of IRE-derived AGO-associated stem-loops could be investigated using several approaches. For instance, one could alter iron concentrations and assay for changes in levels of IRE-derived AGO-associated stem-loops by northern blot. Alternatively, these RNAs could be transfected directly into cells and affects on iron metabolism could be assayed. Given the importance of the IREs to iron metabolism, it will be important to fully understand the function(s) of these elements during iron metabolism. Extrapolation of this function to the broader repertoire of AGO-associated stem-loops derived from mRNAs, will enhance our understanding of the function of these novel elements.

5.4 CONCLUDING REMARKS

RBPs and their RNA targets lie at the heart of post-transcriptional gene regulation. In this dissertation, I have described three projects connected by the common theme of RNA-protein interactions. While these studies are diverse, they each demonstrate that new insights can be gained from global studies of proteins and their RNA target sites. In Chapter 2, we developed a new method to profile RNA-protein interaction sites throughout a transcriptome of interest, laying the groundwork for future studies. In Chapter 3, we identified the genomically encoded target sites of PABPC1, revealing a broader role for this important RBP. Finally, in Chapter 4, we leveraged the interaction of AGO with pre-miRNAs to obtain unprecedented coverage of pre-miRNAs in human and mouse transcriptomes, detailing known miRNAs and identifying novel

cleaved elements in mRNAs with unknown functions. In each of these projects, we have confirmed previous evidence and uncovered novel unexpected findings, opening up new avenues of investigation. Fortunately, these studies detailed molecular maps, which can guide future explorations of these important mediators of post-transcriptional gene regulation.

BIBLIOGRAPHY

1. Crick, F.H., *On protein synthesis.* Symp Soc Exp Biol, 1958. **12**: p. 138-63.

2. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.

3. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

4. Berget, S.M., C. Moore, and P.A. Sharp, *Spliced segments at the 5' terminus of adenovirus 2 late mRNA.* Proc Natl Acad Sci U S A, 1977. **74**(8): p. 3171-5.

5. Merkin, J., C. Russell, P. Chen, and C.B. Burge, *Evolutionary dynamics of gene and isoform regulation in Mammalian tissues.* Science, 2012. **338**(6114): p. 1593-9.

6. Barash, Y., J.A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, . . . B.J. Frey, *Deciphering the splicing code.* Nature, 2010. **465**(7294): p. 53-9.

7. Fu, X.D. and M. Ares, Jr., *Context-dependent control of alternative splicing by RNA-binding proteins.* Nat Rev Genet, 2014. **15**(10): p. 689-701.

8. Elkon, R., A.P. Ugalde, and R. Agami, *Alternative cleavage and polyadenylation: extent, regulation and function.* Nat Rev Genet, 2013. **14**(7): p. 496-506.

9. Mandel, C.R., Y. Bai, and L. Tong, *Protein factors in pre-mRNA 3'-end processing.* Cell Mol Life Sci, 2008. **65**(7-8): p. 1099-122.

10. Luo, W. and D. Bentley, *A ribonucleolytic rat torpedoes RNA polymerase II.* Cell, 2004. **119**(7): p. 911-4.

11. Shi, Y., *Alternative polyadenylation: new insights from global analyses.* RNA, 2012. **18**(12): p. 2105-17.

12. Rabani, M., J.Z. Levin, L. Fan, X. Adiconis, R. Raychowdhury, M. Garber, . . . A. Regev, *Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells.* Nat Biotechnol, 2011. **29**(5): p. 436-42.

13. Ingolia, N.T., S. Ghaemmaghami, J.R. Newman, and J.S. Weissman, *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* Science, 2009. **324**(5924): p. 218-23.

14. Liebhaber, S.A., *mRNA stability and the control of gene expression.* Nucleic Acids Symp Ser, 1997(36): p. 29-32.

15. Russell, J.E., J. Morales, and S.A. Liebhaber, *The role of mRNA stability in the control of globin gene expression.* Prog Nucleic Acid Res Mol Biol, 1997. **57**: p. 249-87.

16. Ameres, S.L. and P.D. Zamore, *Diversifying microRNA sequence and function.* Nat Rev Mol Cell Biol, 2013. **14**(8): p. 475-88.

17. Djuranovic, S., A. Nahvi, and R. Green, *miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay.* Science, 2012. **336**(6078): p. 237-40.

18. Eulalio, A., I. Behm-Ansmant, and E. Izaurralde, *P bodies: at the crossroads of post-transcriptional pathways.* Nat Rev Mol Cell Biol, 2007. **8**(1): p. 9-22.

19. Keene, J.D., *Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome.* Proc Natl Acad Sci U S A, 2001. **98**(13): p. 7018-24.

20. Ascano, M., M. Hafner, P. Cekan, S. Gerstberger, and T. Tuschl, *Identification of RNA-protein interaction networks using PAR-CLIP.* Wiley Interdiscip Rev RNA, 2012. **3**(2): p. 159-77.

21. Keene, J.D., *RNA regulons: coordination of post-transcriptional events.* Nat Rev Genet, 2007. **8**(7): p. 533-43.

22. Baltz, A.G., M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, . . . M. Landthaler, *The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts.* Mol Cell, 2012. **46**: p. 674-90.

23. Castello, A., B. Fischer, K. Eichelbaum, R. Horos, B.M. Beckmann, C. Strein, . . . M.W. Hentze, *Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.* Cell, 2012. **149**: p. 1393-406.

24. Gerstberger, S., M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins.* Nat Rev Genet, 2014. **15**(12): p. 829-45.

25. Glisovic, T., J.L. Bachorik, J. Yong, and G. Dreyfuss, *RNA-binding proteins and post-transcriptional gene regulation.* FEBS letters, 2008. **582**(14): p. 1977-86.

26. Cooper, T.A., L. Wan, and G. Dreyfuss, *RNA and disease.* Cell, 2009. **136**(4): p. 777-93.

27. Masliah, G., P. Barraud, and F.H. Allain, *RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence.* Cell Mol Life Sci, 2013. **70**(11): p. 1875-95.

28. Ambrosone, A., A. Costa, A. Leone, and S. Grillo, *Beyond transcription: RNA-binding proteins as emerging regulators of plant response to environmental constraints.* Plant Sci, 2012. **182**: p. 12-8.

29. Lorkovic, Z.J., *Role of plant RNA-binding proteins in development, stress response and genome organization.* Trends Plant Sci, 2009. **14**(4): p. 229-36.

30. Boucher, L., C.A. Ouzounis, A.J. Enright, and B.J. Blencowe, *A genome-wide survey of RS domain proteins.* RNA, 2001. **7**(12): p. 1693-701.

31. Mousavi, A. and Y. Hotta, *Glycine-rich proteins: a class of novel proteins.* Appl Biochem Biotechnol, 2005. **120**(3): p. 169-74.

32. Fabian, M.R., G. Mathonnet, T. Sundermeier, H. Mathys, J.T. Zipprich, Y.V. Svitkin, . . . N. Sonenberg, *Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation.* Mol Cell, 2009. **35**(6): p. 868-80.

33. Michlewski, G., J.R. Sanford, and J.F. Caceres, *The splicing factor SF2/ASF regulates translation initiation by enhancing phosphorylation of 4E-BP1.* Mol Cell, 2008. **30**(2): p. 179-89.

34. Wu, H., S. Sun, K. Tu, Y. Gao, B. Xie, A.R. Krainer, and J. Zhu, *A splicing-independent function of SF2/ASF in microRNA processing.* Mol Cell, 2010. **38**(1): p. 67-77.

35. Triboulet, R., H.M. Chang, R.J. Lapierre, and R.I. Gregory, *Post-transcriptional control of DGCR8 expression by the Microprocessor.* RNA, 2009. **15**(6): p. 1005-11.

36. Damianov, A. and D.L. Black, *Autoregulation of Fox protein expression to produce dominant negative splicing factors.* RNA, 2010. **16**(2): p. 405-16.

37. Polymenidou, M., C. Lagier-Tourenne, K.R. Hutt, S.C. Huelga, J. Moran, T.Y. Liang, . . . D.W. Cleveland, *Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43.* Nat Neurosci, 2011. **14**(4): p. 459-68.

38. Dai, W., G. Zhang, and E.V. Makeyev, *RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage.* Nucleic Acids Res, 2012. **40**(2): p. 787-800.

39. Hogan, D.J., D.P. Riordan, A.P. Gerber, D. Herschlag, and P.O. Brown, *Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system.* PLoS Biol, 2008. **6**(10): p. e255.

40. Burgess, H.M. and N.K. Gray, *mRNA-specific regulation of translation by poly(A)-binding proteins.* Biochem Soc Trans, 2010. **38**(6): p. 1517-22.

41. Bag, J. and J. Wu, *Translational control of poly(A)-binding protein expression.* Eur J Biochem, 1996. **237**(1): p. 143-52.

42. Mangus, D.A., M.C. Evans, and A. Jacobson, *Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression.* Genome Biol, 2003. **4**(7): p. 223.

43. Wahle, E., *A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation.* Cell, 1991. **66**(4): p. 759-68.

44. Good, P.J., L. Abler, D. Herring, and M.D. Sheets, *Xenopus embryonic poly(A) binding protein 2 (ePABP2) defines a new family of cytoplasmic poly(A) binding proteins expressed during the early stages of vertebrate development.* Genesis, 2004. **38**(4): p. 166-75.

45. Adam, S.A., T. Nakagawa, M.S. Swanson, T.K. Woodruff, and G. Dreyfuss, *mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence.* Mol Cell Biol, 1986. **6**(8): p. 2932-43.

46. Burd, C.G. and G. Dreyfuss, *Conserved structures and diversity of functions of RNA-binding proteins.* Science, 1994. **265**(5172): p. 615-21.

47. Sladic, R.T., C.A. Lagnado, C.J. Bagley, and G.J. Goodall, *Human PABP binds AU-rich RNA via RNA-binding domains 3 and 4.* Eur J Biochem, 2004. **271**(2): p. 450-7.

48. Kleene, K.C., M.Y. Wang, M. Cutler, C. Hall, and D. Shih, *Developmental expression of poly(A) binding protein mRNAs during spermatogenesis in the mouse.* Mol Reprod Dev, 1994. **39**(4): p. 355-64.

49. Blobel, G., *A protein of molecular weight 78,000 bound to the polyadenylate region of eukaryotic messenger RNAs.* Proc Natl Acad Sci U S A, 1973. **70**(3): p. 924-8.

50. Kahvejian, A., Y.V. Svitkin, R. Sukarieh, M.N. M'Boutchou, and N. Sonenberg, *Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms.* Genes Dev, 2005. **19**(1): p. 104-13.

51. Wang, Z. and M. Kiledjian, *The poly(A)-binding protein and an mRNA stability protein jointly regulate an endoribonuclease activity.* Mol Cell Biol, 2000. **20**(17): p. 6334-41.

52. Wilusz, C.J., M. Gao, C.L. Jones, J. Wilusz, and S.W. Peltz, *Poly(A)-binding proteins regulate both mRNA deadenylation and decapping in yeast cytoplasmic extracts.* RNA, 2001. **7**(10): p. 1416-24.

53. Wells, S.E., P.E. Hillner, R.D. Vale, and A.B. Sachs, *Circularization of mRNA by eukaryotic translation initiation factors.* Mol Cell, 1998. **2**(1): p. 135-40.

54. Tarun, S.Z., Jr. and A.B. Sachs, *A common function for mRNA 5' and 3' ends in translation initiation in yeast.* Genes Dev, 1995. **9**(23): p. 2997-3007.

55. Huntzinger, E., J.E. Braun, S. Heimstadt, L. Zekri, and E. Izaurralde, *Two PABPC1-binding sites in GW182 proteins promote miRNA-mediated gene silencing.* EMBO J, 2010. **29**(24): p. 4146-60.

56. de Melo Neto, O.P., N. Standart, and C. Martins de Sa, *Autoregulation of poly(A)-binding protein synthesis in vitro.* Nucleic Acids Res, 1995. **23**(12): p. 2198-205.

57. Hornstein, E., A. Git, I. Braunstein, D. Avni, and O. Meyuhas, *The expression of poly(A)-binding protein gene is translationally regulated in a growth-dependent fashion through a 5'-terminal oligopyrimidine tract motif.* J Biol Chem, 1999. **274**(3): p. 1708-14.

58. Belostotsky, D.A., *Unexpected complexity of poly(A)-binding protein gene families in flowering plants: three conserved lineages that are at least 200 million years old and possible auto- and cross-regulation.* Genetics, 2003. **163**(1): p. 311-9.

59. Baejen, C., P. Torkler, S. Gressel, K. Essig, J. Soding, and P. Cramer, *Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition.* Mol Cell, 2014. **55**(5): p. 745-57.

60. Guo, Z. and F. Sherman, *3'-end-forming signals of yeast mRNA.* Mol Cell Biol, 1995. **15**(11): p. 5983-90.

61. Tuck, A.C. and D. Tollervey, *A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs.* Cell, 2013. **154**(5): p. 996-1009.

62. Miranda, K.C., T. Huynh, Y. Tay, Y.S. Ang, W.L. Tam, A.M. Thomson, . . . I. Rigoutsos, *A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.* Cell, 2006. **126**(6): p. 1203-17.

63. Li, Y. and K.V. Kowdley, *MicroRNAs in common human diseases.* Genomics Proteomics Bioinformatics, 2012. **10**(5): p. 246-53.

64. Altuvia, Y., P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, . . . H. Margalit, *Clustering and conservation patterns of human microRNAs.* Nucleic Acids Res, 2005. **33**(8): p. 2697-706.

65. Ambros, V., *The functions of animal microRNAs.* Nature, 2004. **431**(7006): p. 350-5.

66. Kim, V.N., J. Han, and M.C. Siomi, *Biogenesis of small RNAs in animals.* Nat Rev Mol Cell Biol, 2009. **10**(2): p. 126-39.

67. Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, . . . V.N. Kim, *The nuclear RNase III Drosha initiates microRNA processing.* Nature, 2003. **425**(6956): p. 415-9.

68. Gregory, R.I., K.P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar, *The Microprocessor complex mediates the genesis of microRNAs.* Nature, 2004. **432**(7014): p. 235-40.

69. Heo, I., M. Ha, J. Lim, M.-J. Yoon, J.-E. Park, S.C. Kwon, . . . V.N. Kim, *Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs.* Cell, 2012. **151**(3): p. 521-532.

70. Okamura, K., J.W. Hagen, H. Duan, D.M. Tyler, and E.C. Lai, *The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila.* Cell, 2007. **130**(1): p. 89-100.

71. Ruby, J.G., C.H. Jan, and D.P. Bartel, *Intronic microRNA precursors that bypass Drosha processing.* Nature, 2007. **448**(7149): p. 83-6.

72. Yi, R., Y. Qin, I.G. Macara, and B.R. Cullen, *Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs.* Genes Dev, 2003. **17**(24): p. 3011-6.

73. Lund, E., S. Guttinger, A. Calado, J.E. Dahlberg, and U. Kutay, *Nuclear export of microRNA precursors.* Science, 2004. **303**(5654): p. 95-8.

74. Chendrimada, T.P., R.I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar, *TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing.* Nature, 2005. **436**(7051): p. 740-4.

75. Gregory, R.I., T.P. Chendrimada, N. Cooch, and R. Shiekhattar, *Human RISC couples microRNA biogenesis and posttranscriptional gene silencing.* Cell, 2005. **123**(4): p. 631-40.

76. Maniataki, E. and Z. Mourelatos, *A human, ATP-independent, RISC assembly machine fueled by pre-miRNA.* Genes Dev, 2005. **19**(24): p. 2979-90.

77. Liu, X., D.-Y. Jin, M.T. McManus, and Z. Mourelatos, *Precursor microRNA-programmed silencing complex assembly pathways in mammals.* Molecular Cell, 2012. **46**(4): p. 507-517.

78. Diederichs, S. and D.A. Haber, *Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression.* Cell, 2007. **131**(6): p. 1097-108.

79. Cheloufi, S., C.O. Dos Santos, M.M. Chong, and G.J. Hannon, *A dicer-independent miRNA biogenesis pathway that requires Ago catalysis.* Nature, 2010. **465**(7298): p. 584-9.

80. Cifuentes, D., H. Xue, D.W. Taylor, H. Patnode, Y. Mishima, S. Cheloufi, . . . A.J. Giraldez, *A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity.* Science, 2010. **328**(5986): p. 1694-8.

81. Bang, C., S. Batkai, S. Dangwal, S.K. Gupta, A. Foinquinos, A. Holzmann, . . . T. Thum, *Cardiac fibroblast-derived microRNA passenger strand-enriched exosomes mediate cardiomyocyte hypertrophy.* J Clin Invest, 2014. **124**(5): p. 2136-46.

82. Li, S.C., Y.L. Liao, M.R. Ho, K.W. Tsai, C.H. Lai, and W.C. Lin, *miRNA arm selection and isomiR distribution in gastric cancer.* BMC Genomics, 2012. **13 Suppl 1**: p. S13.

83. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.

84. Helwak, A., G. Kudla, T. Dudnakova, and D. Tollervey, *Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding.* Cell, 2013. **153**(3): p. 654-65.

85. Grimson, A., K.K. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel, *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.

86. Wu, L., J. Fan, and J.G. Belasco, *MicroRNAs direct rapid deadenylation of mRNA.* Proc Natl Acad Sci U S A, 2006. **103**(11): p. 4034-9.

87. Guo, H., N.T. Ingolia, J.S. Weissman, and D.P. Bartel, *Mammalian microRNAs predominantly act to decrease target mRNA levels.* Nature, 2010. **466**(7308): p. 835-40.

88. Bazzini, A.A., M.T. Lee, and A.J. Giraldez, *Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish.* Science, 2012. **336**(6078): p. 233-7.

89. Macias, S., M. Plass, A. Stajuda, G. Michlewski, E. Eyras, and J.F. Caceres, *DGCR8 HITS-CLIP reveals novel functions for the Microprocessor.* Nat Struct Mol Biol, 2012. **19**(8): p. 760-6.

90. Allerson, C.R., A. Martinez, E. Yikilmaz, and T.A. Rouault, *A high-capacity RNA affinity column for the purification of human IRP1 and IRP2 overexpressed in Pichia pastoris.* RNA, 2003. **9**(3): p. 364-74.

91. Choi, Y.D. and G. Dreyfuss, *Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): a unique supramolecular assembly.* Proc Natl Acad Sci U S A, 1984. **81**(23): p. 7471-5.

92. Hellman, L.M. and M.G. Fried, *Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions.* Nat Protoc, 2007. **2**(8): p. 1849-61.

93. Kenan, D.J. and J.D. Keene, *In vitro selection of aptamers from RNA libraries.* Methods Mol Biol, 1999. **118**: p. 217-31.

94. Ray, D., H. Kazan, E.T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, . . . T.R. Hughes, *Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins.* Nat Biotechnol, 2009. **27**(7): p. 667-70.

95. Ray, D., H. Kazan, K.B. Cook, M.T. Weirauch, H.S. Najafabadi, X. Li, . . . T.R. Hughes, *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.

96. Barkan, A., *Genome-wide analysis of RNA-protein interactions in plants.* Methods Mol Biol, 2009. **553**: p. 13-37.

97. Selth, L.A., C. Gilbert, and J.Q. Svejstrup, *RNA immunoprecipitation to determine RNA-protein associations in vivo.* Cold Spring Harb Protoc, 2009. **2009**(6): p. pdb prot5234.

98. Ule, J., K. Jensen, A. Mele, and R.B. Darnell, *CLIP: a method for identifying protein-RNA interaction sites in living cells.* Methods, 2005. **37**(4): p. 376-86.

99. Hafner, M., S. Lianoglou, T. Tuschl, and D. Betel, *Genome-wide identification of miRNA targets by PAR-CLIP.* Methods, 2012.

100. Konig, J., K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, . . . J. Ule, *iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution.* J Vis Exp, 2011(50).

101. Ule, J., K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R.B. Darnell, *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 2003. **302**(5648): p. 1212-5.

102. Huelga, S.C., A.Q. Vu, J.D. Arnold, T.Y. Liang, P.P. Liu, B.Y. Yan, . . . G.W. Yeo, *Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins.* Cell Reports, 2012. **1**(2): p. 167-78.

103. Sanford, J.R., X. Wang, M. Mort, N. Vanduyn, D.N. Cooper, S.D. Mooney, . . . Y. Liu, *Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts.* Genome Res, 2009. **19**(3): p. 381-94.

104. Goodarzi, H., H.S. Najafabadi, P. Oikonomou, T.M. Greco, L. Fish, R. Salavati, . . . S. Tavazoie, *Systematic discovery of structural elements governing stability of mammalian messenger RNAs.* Nature, 2012. **485**(7397): p. 264-8.

105. Darnell, J.C., S.J. Van Driesche, C. Zhang, K.Y. Hung, A. Mele, C.E. Fraser, . . . R.B. Darnell, *FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism.* Cell, 2011. **146**(2): p. 247-61.

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

107. Schueler, M., M. Munschauer, L.H. Gregersen, A. Finzel, A. Loewer, W. Chen, . . . C. Dieterich, *Differential protein occupancy profiling of the mRNA transcriptome.* Genome Biol, 2014. **15**(1): p. R15.

108. Gosai, S.J., S.W. Foley, D. Wang, I.M. Silverman, N. Selamoglu, A.D. Nelson, . . . B.D. Gregory, *Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus.* Mol Cell, 2015. **57**(2): p. 376-88.

109. Rinn, J.L., M. Kertesz, J.K. Wang, S.L. Squazzo, X. Xu, S.A. Brugmann, . . . H.Y. Chang, *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.

110. Guttman, M., J. Donaghey, B.W. Carey, M. Garber, J.K. Grenier, G. Munson, . . . E.S. Lander, *lincRNAs act in the circuitry controlling pluripotency and differentiation.* Nature, 2011. **477**(7364): p. 295-300.

111. Dreyfuss, G., M.J. Matunis, S. Pinol-Roma, and C.G. Burd, *hnRNP proteins and the biogenesis of mRNA.* Annu Rev Biochem, 1993. **62**: p. 289-321.

112. Glisovic, T., J.L. Bachorik, J. Yong, and G. Dreyfuss, *RNA-binding proteins and post-transcriptional gene regulation.* FEBS Lett, 2008. **582**(14): p. 1977-86.

113. Keene, J.D. and S.A. Tenenbaum, *Eukaryotic mRNPs may represent posttranscriptional operons.* Mol Cell, 2002. **9**(6): p. 1161-7.

114. Aerts, S., *Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets.* Curr Top Dev Biol, 2012. **98**: p. 121-45.

115. Konig, J., K. Zarnack, N.M. Luscombe, and J. Ule, *Protein-RNA interactions: new genomic technologies and perspectives.* Nat Rev Genet, 2011. **13**(2): p. 77-83.

116. Hafner, M., M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, . . . T. Tuschl, *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.

117. Hoell, J.I., E. Larsson, S. Runge, J.D. Nusbaum, S. Duggimpudi, T.A. Farazi, . . . T. Tuschl, *RNA targets of wild-type and mutant FET family proteins.* Nat Struct Mol Biol, 2011. **18**(12): p. 1428-31.

118. Huelga, S.C., A.Q. Vu, J.D. Arnold, T.Y. Liang, P.P. Liu, B.Y. Yan, . . . G.W. Yeo, *Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins.* Cell Rep, 2012. **1**(2): p. 167-78.

119. Konig, J., K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, . . . J. Ule, *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.* Nat Struct Mol Biol, 2010. **17**(7): p. 909-15.

120. Lebedeva, S., M. Jens, K. Theil, B. Schwanhausser, M. Selbach, M. Landthaler, and N. Rajewsky, *Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR.* Mol Cell, 2011. **43**(3): p. 340-52.

121. Mukherjee, N., D.L. Corcoran, J.D. Nusbaum, D.W. Reid, S. Georgiev, M. Hafner, . . . J.D. Keene, *Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability.* Mol Cell, 2011. **43**(3): p. 327-39.

122. Xue, Y., Y. Zhou, T. Wu, T. Zhu, X. Ji, Y.S. Kwon, . . . Y. Zhang, *Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping.* Mol Cell, 2009. **36**(6): p. 996-1006.

123. Wang, Z., M. Kayikci, M. Briese, K. Zarnack, N.M. Luscombe, G. Rot, . . . J. Ule, *iCLIP predicts the dual splicing effects of TIA-RNA interactions.* PLoS Biol, 2010. **8**(10): p. e1000530.

124. Wilbert, M.L., S.C. Huelga, K. Kapeli, T.J. Stark, T.Y. Liang, S.X. Chen, . . . G.W. Yeo, *LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance.* Mol Cell, 2012. **48**(2): p. 195-206.

125. Kishore, S., L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, *A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.* Nat Methods, 2011. **8**(7): p. 559-64.

126. Sievers, C., T. Schlumpf, R. Sawarkar, F. Comoglio, and R. Paro, *Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data.* Nucleic Acids Res, 2012. **40**(20): p. e160.

127. Baltz, A.G., M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, . . . M. Landthaler, *The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts.* Mol Cell, 2012. **46**(5): p. 674-90.

128. Freeberg, M.A., T. Han, J.J. Moresco, A. Kong, Y.C. Yang, Z.J. Lu, . . . J.K. Kim, *Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae.* Genome Biol, 2013. **14**(2): p. R13.

129. Li, F., Q. Zheng, P. Ryvkin, I. Dragomir, Y. Desai, S. Aiyer, . . . B.D. Gregory, *Global analysis of RNA secondary structure in two metazoans.* Cell Rep, 2012. **1**(1): p. 69-82.

130. Zheng, Q., P. Ryvkin, F. Li, I. Dragomir, O. Valladares, J. Yang, . . . B.D. Gregory, *Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis.* PLoS Genet, 2010. **6**(9): p. e1001141.

131. Jensen, K.B. and R.B. Darnell, *CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins.* Methods Mol Biol, 2008. **488**: p. 85-98.

132. Tenenbaum, S.A., P.J. Lager, C.C. Carson, and J.D. Keene, *Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays.* Methods, 2002. **26**(2): p. 191-8.

133. Gilbert, C. and J.Q. Svejstrup, *RNA immunoprecipitation for determining RNA-protein associations in vivo.* Curr Protoc Mol Biol, 2006. **Chapter 27**: p. Unit 27 4.

134. Muino, J.M., K. Kaufmann, R.C. van Ham, G.C. Angenent, and P. Krajewski, *ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions.* Plant Methods, 2011. **7**: p. 11.

135. Bailey, T.L., M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, . . . W.S. Noble, *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.

136. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics, 2011. **27**(7): p. 1017-8.

137. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

138. Capon, F., M.H. Allen, M. Ameen, A.D. Burden, D. Tillman, J.N. Barker, and R.C. Trembath, *A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups.* Hum Mol Genet, 2004. **13**(20): p. 2361-8.

139. Chamary, J.V., J.L. Parmley, and L.D. Hurst, *Hearing silence: non-neutral evolution at synonymous sites in mammals.* Nat Rev Genet, 2006. **7**(2): p. 98-108.

140. Brest, P., P. Lapaquette, M. Souidi, K. Lebrigand, A. Cesaro, V. Vouret-Craviari, . . . P. Hofman, *A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease.* Nat Genet, 2011. **43**(3): p. 242-5.

141. Sauna, Z.E. and C. Kimchi-Sarfaty, *Understanding the contribution of synonymous mutations to human disease.* Nat Rev Genet, 2011. **12**(10): p. 683-91.

142. Hindorff, L.A., P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, and T.A. Manolio, *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

143. Habegger, L., A. Sboner, T.A. Gianoulis, J. Rozowsky, A. Agarwal, M. Snyder, and M. Gerstein, *RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.* Bioinformatics, 2011. **27**(2): p. 281-3.

144. Garber, M., M. Guttman, M. Clamp, M.C. Zody, N. Friedman, and X. Xie, *Identifying novel constrained elements by exploiting biased substitution patterns.* Bioinformatics, 2009. **25**(12): p. i54-62.

145. Goss, D.J. and F.E. Kleiman, *Poly(A) binding proteins: are they all created equal?* Wiley Interdiscip Rev RNA, 2013. **4**(2): p. 167-79.

146. Zhang, C. and R.B. Darnell, *Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data.* Nat Biotechnol, 2011. **29**(7): p. 607-14.

147. Moore, M.J., C. Zhang, E.C. Gantman, A. Mele, J.C. Darnell, and R.B. Darnell, *Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis.* Nat Protoc, 2014. **9**(2): p. 263-93.

148. Williams, K.R. and W.H. Konigsberg, *Identification of amino acid residues at interface of protein-nucleic acid complexes by photochemical cross-linking.* Methods Enzymol, 1991. **208**: p. 516-39.

149. Sugimoto, Y., J. Konig, S. Hussain, B. Zupan, T. Curk, M. Frye, and J. Ule, *Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.* Genome Biol, 2012. **13**(8): p. R67.

150. Tian, B., J. Hu, H. Zhang, and C.S. Lutz, *A large-scale analysis of mRNA polyadenylation of human and mouse genes.* Nucleic Acids Res, 2005. **33**(1): p. 201-12.

151. Althammer, S., J. Gonzalez-Vallinas, C. Ballare, M. Beato, and E. Eyras, *Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data.* Bioinformatics, 2011. **27**(24): p. 3333-40.

152. Pesole, G., S. Liuni, G. Grillo, and C. Saccone, *Structural and compositional features of untranslated regions of eukaryotic mRNAs.* Gene, 1997. **205**(1-2): p. 95-102.

153. Melo, E.O., O.P. de Melo Neto, and C. Martins de Sa, *Adenosine-rich elements present in the 5'-untranslated region of PABP mRNA can selectively reduce the abundance and translation of CAT mRNAs in vivo.* FEBS Lett, 2003. **546**(2-3): p. 329-34.

154. Melo, E.O., R. Dhalia, C. Martins de Sa, N. Standart, and O.P. de Melo Neto, *Identification of a C-terminal poly(A)-binding protein (PABP)-PABP interaction domain: role in cooperative binding to poly (A) and efficient cap distal translational repression.* J Biol Chem, 2003. **278**(47): p. 46357-68.

155. Nayler, O., W. Stratling, J.P. Bourquin, I. Stagljar, L. Lindemann, H. Jasper, . . . S. Stamm, *SAF-B protein couples transcription and pre-mRNA splicing to SAR/MAR elements.* Nucleic Acids Res, 1998. **26**(15): p. 3542-9.

156. Paasinen-Sohns, A., E. Kaariainen, M. Yin, K. Jarvinen, P. Nummela, and E. Holtta, *Chaotic neovascularization induced by aggressive fibrosarcoma cells overexpressing S-adenosylmethionine decarboxylase.* Int J Biochem Cell Biol, 2011. **43**(3): p. 441-54.

157. Tabassum, R., A. Jaiswal, G. Chauhan, O.P. Dwivedi, S. Ghosh, R.K. Marwaha, . . . D. Bharadwaj, *Genetic variant of AMD1 is associated with obesity in urban Indian children.* PLoS One, 2012. **7**(4): p. e33162.

158. de Melo Neto, O.P., J.A. Walker, C.M. Martins de Sa, and N. Standart, *Levels of free PABP are limited by newly polyadenylated mRNA in early Spisula embryogenesis.* Nucleic Acids Res, 2000. **28**(17): p. 3346-53.

159. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.

160. Smith, R.W., T.K. Blee, and N.K. Gray, *Poly(A)-binding proteins are required for diverse biological processes in metazoans.* Biochem Soc Trans, 2014. **42**(4): p. 1229-37.

161. Sachs, A.B., R.W. Davis, and R.D. Kornberg, *A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability.* Mol Cell Biol, 1987. **7**(9): p. 3268-76.

162. Sigrist, S.J., P.R. Thiel, D.F. Reiff, P.E. Lachance, P. Lasko, and C.M. Schuster, *Postsynaptic translation affects the efficacy and morphology of neuromuscular junctions.* Nature, 2000. **405**(6790): p. 1062-5.

163. de Moor, C.H., H. Meijer, and S. Lissenden, *Mechanisms of translational control by the 3' UTR in development and differentiation.* Semin Cell Dev Biol, 2005. **16**(1): p. 49-58.

164. Gebauer, F., T. Preiss, and M.W. Hentze, *From cis-regulatory elements to complex RNPs and back.* Cold Spring Harb Perspect Biol, 2012. **4**(7): p. a012245.

165. Richter, J.D., *CPEB: a life in translation.* Trends Biochem Sci, 2007. **32**(6): p. 279-85.

166. Theis, M., K. Si, and E.R. Kandel, *Two previously undescribed members of the mouse CPEB family of genes and their inducible expression in the principal cell layers of the hippocampus.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9602-7.

167. Uchida, N., S. Hoshino, H. Imataka, N. Sonenberg, and T. Katada, *A novel role of the mammalian GSPT/eRF3 associating with poly(A)-binding protein in Cap/Poly(A)-dependent translation.* J Biol Chem, 2002. **277**(52): p. 50286-92.

168. Peixeiro, I., A. Inacio, C. Barbosa, A.L. Silva, S.A. Liebhaber, and L. Romao, *Interaction of PABPC1 with the translation initiation complex is critical to the NMD resistance of AUG-proximal nonsense mutations.* Nucleic Acids Res, 2012. **40**(3): p. 1160-73.

169. Yanagiya, A., G. Delbes, Y.V. Svitkin, B. Robaire, and N. Sonenberg, *The poly(A)-binding protein partner Paip2a controls translation during late spermiogenesis in mice.* J Clin Invest, 2010. **120**(9): p. 3389-400.

170. Verlaet, M., V. Deregowski, G. Denis, C. Humblet, M.T. Stalmans, V. Bours, . . . M.P. Defresne, *Genetic imbalances in preleukemic thymuses.* Biochem Biophys Res Commun, 2001. **283**(1): p. 12-8.

171. Chi, S.W., J.B. Zang, A. Mele, and R.B. Darnell, *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.

172. Elliott, R., F. Li, I. Dragomir, M.M. Chua, B.D. Gregory, and S.R. Weiss, *Analysis of the host transcriptome from demyelinating spinal cord of murine coronavirus-infected mice.* PLoS One, 2013. **8**(9): p. e75346.

173. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

174. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

175. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**: p. 10-12.

176. Bembom, O., S. Keles, and M.J. van der Laan, *Supervised detection of conserved motifs in DNA sequences with cosmo.* Stat Appl Genet Mol Biol, 2007. **6**: p. Article8.

177. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

178. Ingolia, N.T., G.A. Brar, S. Rouskin, A.M. McGeachy, and J.S. Weissman, *The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments.* Nat Protoc, 2012. **7**(8): p. 1534-50.

179. Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S.L. Salzberg, *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.

180. Humphreys, D.T., B.J. Westman, D.I. Martin, and T. Preiss, *MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function.* Proc Natl Acad Sci U S A, 2005. **102**(47): p. 16961-6.

181. Yoda, M., D. Cifuentes, N. Izumi, Y. Sakaguchi, T. Suzuki, A.J. Giraldez, and Y. Tomari, *Poly(A)-specific ribonuclease mediates 3'-end trimming of Argonaute2-cleaved precursor microRNAs.* Cell Rep, 2013. **5**(3): p. 715-26.

182. Heo, I., C. Joo, J. Cho, M. Ha, J. Han, and V.N. Kim, *Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA.* Mol Cell, 2008. **32**(2): p. 276-84.

183. Liu, X., Q. Zheng, N. Vrettos, M. Maragkakis, P. Alexiou, B.D. Gregory, and Z. Mourelatos, *A MicroRNA precursor surveillance system in quality control of MicroRNA synthesis.* Molecular Cell, 2014. **55**(6): p. 868-879.

184. Kim, B., M. Ha, L. Loeff, H. Chang, D.K. Simanshu, S. Li, . . . V.N. Kim, *TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms.* The EMBO journal, 2015. **34**(13): p. 1801-1815.

185. Newman, M.A., V. Mani, and S.M. Hammond, *Deep sequencing of microRNA precursors reveals extensive 3' end modification.* RNA (New York, N.Y.), 2011. **17**(10): p. 1795-1803.

186. Friedlander, M.R., W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky, *Discovering microRNAs from deep sequencing data using miRDeep.* Nat Biotechnol, 2008. **26**(4): p. 407-15.

187. Londin, E., P. Loher, A.G. Telonis, K. Quann, P. Clark, Y. Jing, . . . I. Rigoutsos, *Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs.* Proc Natl Acad Sci U S A, 2015. **112**(10): p. E1106-15.

188. Li, N., X. You, T. Chen, S.D. Mackowiak, M.R. Friedländer, M. Weigt, . . . W. Chen, *Global profiling of miRNAs and the hairpin precursors: insights into miRNA processing and novel miRNA discovery.* Nucleic Acids Research, 2013. **41**(6): p. 3619-3634.

189. Burroughs, A.M., M. Kawano, Y. Ando, C.O. Daub, and Y. Hayashizaki, *pre-miRNA profiles obtained through application of locked nucleic acids and deep sequencing reveals complex 5′/3' arm variation including concomitant cleavage and polyuridylation patterns.* Nucleic Acids Research, 2012. **40**(4): p. 1424-1437.

190. Nelson, P.T., M. De Planell-Saguer, S. Lamprinaki, M. Kiriakidou, P. Zhang, U. O'Doherty, and Z. Mourelatos, *A novel monoclonal antibody against human Argonaute proteins reveals unexpected characteristics of miRNAs in human blood cells.* RNA, 2007. **13**(10): p. 1787-92.

191. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

192. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

193. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D68-73.

194. Heo, I., C. Joo, Y.K. Kim, M. Ha, M.J. Yoon, J. Cho, . . . V.N. Kim, *TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation.* Cell, 2009. **138**(4): p. 696-708.

195. Thornton, J.E., P. Du, L. Jing, L. Sjekloca, S. Lin, E. Grossi, . . . R.I. Gregory, *Selective microRNA uridylation by Zcchc6 (TUT7) and Zcchc11 (TUT4).* Nucleic Acids Res, 2014. **42**(18): p. 11777-91.

196. Liu, J., M.A. Carmell, F.V. Rivas, C.G. Marsden, J.M. Thomson, J.J. Song, . . . G.J. Hannon, *Argonaute2 is the catalytic engine of mammalian RNAi.* Science, 2004. **305**(5689): p. 1437-41.

197. Meister, G., M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl, *Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs.* Mol Cell, 2004. **15**(2): p. 185-97.

198. Yekta, S., I.H. Shih, and D.P. Bartel, *MicroRNA-directed cleavage of HOXB8 mRNA.* Science, 2004. **304**(5670): p. 594-6.

199. Han, J., J.S. Pedersen, S.C. Kwon, C.D. Belair, Y.K. Kim, K.H. Yeom, . . . V.N. Kim, *Posttranscriptional crossregulation between Drosha and DGCR8.* Cell, 2009. **136**(1): p. 75-84.

200. Kadener, S., J. Rodriguez, K.C. Abruzzi, Y.L. Khodor, K. Sugino, M.T. Marr, 2nd, . . . M. Rosbash, *Genome-wide identification of targets of the drosha-pasha/DGCR8 complex.* RNA, 2009. **15**(4): p. 537-45.

201. Lorenz, R., S.H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, and I.L. Hofacker, *ViennaRNA Package 2.0.* Algorithms Mol Biol, 2011. **6**: p. 26.

202. Rybak-Wolf, A., M. Jens, Y. Murakawa, M. Herzog, M. Landthaler, and N. Rajewsky, *A variety of dicer substrates in human and C. elegans.* Cell, 2014. **159**(5): p. 1153-67.

203. Spitale, R.C., R.A. Flynn, Q.C. Zhang, P. Crisalli, B. Lee, J.W. Jung, . . . H.Y. Chang, *Structural imprints in vivo decode RNA regulatory mechanisms.* Nature, 2015. **519**(7544): p. 486-90.

204. Hentze, M.W., T.A. Rouault, S.W. Caughman, A. Dancis, J.B. Harford, and R.D. Klausner, *A cis-acting element is necessary and sufficient for translational regulation of human ferritin expression in response to iron.* Proc Natl Acad Sci U S A, 1987. **84**(19): p. 6730-4.

205. Durbin, B.P., J.S. Hardin, D.M. Hawkins, and D.M. Rocke, *A variance-stabilizing transformation for gene-expression microarray data.* Bioinformatics, 2002. **18 Suppl 1**: p. S105-10.

206. Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, *Variance stabilization applied to microarray data calibration and to the quantification of differential expression.* Bioinformatics, 2002. **18 Suppl 1**: p. S96-104.

207. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

208. Silverman, I.M., F. Li, A. Alexander, L. Goff, C. Trapnell, J.L. Rinn, and B.D. Gregory, *RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome.* Genome Biol, 2014. **15**(1): p. R3.

209. Wang, K.C. and H.Y. Chang, *Molecular mechanisms of long noncoding RNAs.* Mol Cell, 2011. **43**(6): p. 904-14.

210. Keene, J.D., *RNA regulons: coordination of post-transcriptional events.* Nat Rev Genet, 2007. **8**: p. 533-43.

211. Keene, J.D. and S.A. Tenenbaum, *Eukaryotic mRNPs may represent posttranscriptional operons.* Mol Cell, 2002. **9**: p. 1161-7.

212. Silverman, I.M. and B.D. Gregory, *Transcriptome-wide ribonuclease-mediated protein footprinting to identify RNA-protein interaction sites.* Methods, 2015. **72**: p. 76-85.

213. Migliaccio, A.R., *Erythroblast enucleation.* Haematologica, 2010. **95**(12): p. 1985-8.

214. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.

215. Kiledjian, M., X. Wang, and S.A. Liebhaber, *Identification of two KH domain proteins in the alpha-globin mRNP stability complex.* EMBO J, 1995. **14**(17): p. 4357-64.

216. Proudfoot, N.J., *Ending the message: poly(A) signals then and now.* Genes Dev, 2011. **25**(17): p. 1770-82.

217. Weill, L., E. Belloc, F.A. Bava, and R. Mendez, *Translational control by changes in poly(A) tail length: recycling mRNAs.* Nat Struct Mol Biol, 2012. **19**(6): p. 577-85.

218. Wakiyama, M., H. Imataka, and N. Sonenberg, *Interaction of eIF4G with poly(A)-binding protein stimulates translation and is critical for Xenopus oocyte maturation.* Curr Biol, 2000. **10**(18): p. 1147-50.

219. Kini, H.K., J. Kong, and S.A. Liebhaber, *Cytoplasmic poly(A) binding protein C4 serves a critical role in erythroid differentiation.* Mol Cell Biol, 2014. **34**(7): p. 1300-9.

220. Reimao-Pinto, M.M., V. Ignatova, T.R. Burkard, J.H. Hung, R.A. Manzenreither, I. Sowemimo, . . . S.L. Ameres, *Uridylation of RNA Hairpins by Tailor Confines the Emergence of MicroRNAs in Drosophila.* Mol Cell, 2015. **59**(2): p. 203-16.

221. Liu, N., M. Abe, L.R. Sabin, G.J. Hendriks, A.S. Naqvi, Z. Yu, . . . N.M. Bonini, *The exoribonuclease Nibbler controls 3' end processing of microRNAs in Drosophila.* Curr Biol, 2011. **21**(22): p. 1888-93.

222. Sen, G.L. and H.M. Blau, *Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies.* Nat Cell Biol, 2005. **7**(6): p. 633-6.

223. Gray, N.K. and M.W. Hentze, *Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs.* EMBO J, 1994. **13**(16): p. 3882-91.