The Development and Application of a Method to Quantitatively Identify RNA Binding Sites, and Whole Transcript Targets of RNA Binding Proteins

by

Cindo O. Nicholson

Department of Molecular Genetics & Microbiology
Duke University

Date:_____

Approved:

_____
Jack D. Keene, Supervisor

_____
Micah Luftig

_____
David Pickup

_____
Christopher Nicchitta

_____
Gerard Blobe

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Molecular Genetics & Microbiology in the Graduate School
of Duke University

2016

ABSTRACT

The Development and Application of a Method to Quantitatively Identify RNA Binding Sites, and Whole Transcript Targets of RNA Binding Proteins

by

Cindo O. Nicholson

Department of Molecular Genetics & Microbiology
Duke University

Date:_____

Approved:

_____
Jack D. Keene, Supervisor

_____
Micah Luftig

_____
David Pickup

_____
Christopher Nicchitta

_____
Gerard Blobe

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Molecular Genetics & Microbiology in the Graduate School of
Duke University

2016

Abstract

RNA binding proteins (RBPs) and non-coding RNAs orchestrate gene expression in part through the recognition of specific sites in mRNA. Thus understanding the connection between binding to specific sites and regulation of the whole transcript is essential. Current methods to do this can either identify the binding sites or quantitate binding to whole transcripts, but not both. Furthermore reliance of binding site detection on ultraviolet crosslinking results in inefficient identification of binding sites, and insufficient data to assess binding strength at sites. I have overcome these limitations by combining aspects of current methods to develop a new method called DO-RIP-seq (digestion optimization RNA immunoprecipitations with deep sequencing) that can quantitate the binding strength of RBPs at sites in mRNA, and also relate binding sites to binding of the whole mRNA. DO-RIP-seq was developed using the well-studied RBP ELAVL1/HuR as a test case, and applied to the less well-studied RBP known as RBM38/RNPC1. The quantitative data from DO-RIP-seq out-performed current binding site methods at predicting other features of the binding sites of HuR and RBM38, for example the lack of RNA secondary structure, and preferences in binding to particular sub-motifs. My studies indicate that DO-RIP-seq will be useful in uncovering the determinants of RNA-protein interactions, and studying dynamic biological processes that could modulate these interactions.

Dedication

I dedicate this dissertation to the memory of my father Cindo O. Nicholson (Sr.) and my father-in-law Michael A. Peart who were both laid to rest during my graduate fellowship. Dad, I thank you for being the best father I could be blessed to have. You were the wisest man I ever knew and I am forever grateful for the many lessons, words of wisdom and encouragement, and of course your love and patience. It is because of your many lessons, and prayers that I have made it this far. Father Peart, I am forever grateful to you for raising an amazing lady who I am most blessed to be married to. I enjoyed and look back fondly on the much too short time we had together. I would like to thank you both for being exemplary men in my life, and I hope that Christine and I have made you both very proud.

Love you always, and forever in my memories.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge all members of the Keene Lab, both former and current, and all the members of my dissertation committee. A special thank you to Jack Keene for providing the opportunity to complete my doctoral dissertation under his supervision. Thank you Jack for being a patient mentor, and for being supportive of my endeavors. Also, I would like to thank Matthew Friedersdorf for working closely with me on my dissertation project. Matt has helped me appreciate that there is more to computers than "point and click". I would like to thank my family and friends who have always supported me with words of encouragement and through prayers.

Finally, I would like to thank my dear, and lovely wife Christine Maria. Christine, you have stuck by me through this roller coaster ride, and I am grateful for your never-ending support and belief in me. You were always cheering the loudest for my accomplishments, and you were the first to help me recover from the numerous days when things did not turn out the way I wanted. You have had to sacrifice so much in order to be my emotional support through my time in graduate school. I love you so very much Christine, and may we continue to move forward.

1. Introduction

The life cycle of mRNA; from its synthesis to its degradation is orchestrated through interactions with RNA-binding proteins (RBPs) and non-coding RNA. During its transcription the pre-mRNA is capped at its 5′ end with a 7-methylguanosine, exons are spliced after removing introns, and the 3′ end is enzymatically cleaved and polyadenylated **[1, 2]**. These processes, which are all orchestrated through interactions with RBPs result in the production of a mature mRNA. The mature mRNA is exported from the nucleus to the cytoplasm by interactions between more RBPs and the nuclear pore complexes **[3, 4]**. In the cytoplasm the mRNA is translated to protein when bound by translation initiation factors and ribosome complexes **[5, 6]**, or can be degraded by exonuclease complexes from the 5′ or 3′ end **[7-9]**. Thus the life cycle of mRNA involves dynamic interactions between various RBPs and non-coding RNAs and ultimately results in the production of proteins needed for various cell functions.

In addition, each post-transcriptional step allows gene expression to be regulated at multiple steps to ensure proper temporal and spatial expression of proteins. This is important for eukaryotic cells when responding to stimuli through alterations in the expression of functionally relevant gene products. At the post-transcriptional level, RBPs tend to target functionally related groups of mRNAs in response to stimuli. The coordinated post-transcriptional regulation of functionally related groups of mRNAs

can be thought of as post-transcriptional RNA regulons **[10-13]**. This is a concept related to but distinct from mechanisms used by prokaryotes. Prokaryotic genomes are polycistronic, i.e. they organize multiple functionally related protein-coding genes under the control off a single promoter that is often responsive to changes in the environment. These are referred to as DNA/transcriptional operons and was first defined by Jacob and colleagues **[14]**. This is advantageous to prokaryotes because transcription is physically coupled to translation. However, eukaryotic genomes largely do not have polycistronic genes, the nascent mRNA requires processing as detailed earlier, and translation of the mRNA is physically non-coupled, as it requires the mRNA to be exported from the membrane-bound nucleus to the cytoplasm to be engaged by ribosomes. Thus, mechanisms to coordinate the expression of functionally related groups of mRNAs by RBPs better suits the genome architecture and sub-cellular compartmentalization observed in eukaryotic cells.

The dynamic life cycle of the mRNA and the coordinated regulation of mRNA subsets also presented questions on mechanisms of how RBPs and non-coding RNAs coordinated these processes. What are, and how do combinations of RNA sequence elements and RBPs that result in cumulative functional interactions? What are the events that govern the assembly of ribonucleoprotein complexes? What are the protein-RNA interactions required for protein synthesis in response to stimuli? To address these

questions required the development of methods that could identify *en masse* the specific determinants of RBP binding, and ultimately identify mRNAs bound and regulated by the RBP. These methods can be placed into two categories: immunoprecipitation of native RNP complexes followed by sequencing, and the immunoprecipitation of crosslinked, RNase treated RNP complexes followed by sequencing.

1.1 Decoding protein-RNA interactions: transcriptome-wide identification of whole transcript targets of RBPs

Gene expression changes at the transcriptional level have been used to infer changes in the proteome of a cell in response to stimuli such as growth/differentiation signals. However, the processes of mRNA de/stabilization and translation are regulated by RBPs, and affect the amount of protein synthesized. As a result, steady-state mRNA abundance may not directly correlate with protein abundance **[15-19]**. Techniques that could identify mRNAs bound by RBPs *en masse* when cells were stimulated were needed. To address this need, Tenenbaum et al. developed a method to isolate mRNAs from immunoprecipitated messenger ribonucleoprotein complexes (mRNPs). The mRNAs were then identified using cDNA arrays **[20]**. Using this technique Tenenbaum et al. showed that the RBPs HuB, PABP, and EIF4E bound different subsets of mRNAs, and the bound mRNA profiles differed from the total mRNA profile of the P19 neuronal cell line. In addition, they showed that the subset of HuB-bound mRNAs changed upon

differentiating the P19 cells by treating them with retinoic acid. Altogether these experiments showed that mRNPs were dynamic and responsive to the cell's state, and indicated that the activity of the RBPs on mRNA were integral to understanding functional outcomes. Later this method was ported to the use of microarrays and referred to as RNA immunoprecipitation and chip hybridization or RIP-chip **[21]**, and then more recently to deep sequencing and referred to as RIP-seq **[22, 23]**.

The utility of RIP-chip for studying mRNP remodeling across multiple physiological states was extended by probabilistically modeling the association of mRNAs with an RBP during changing conditions. Such measurements were possible because of negative control/background RIPs performed in parallel. The benefits of the negative control were two-fold: 1) it accounts for changes in mRNA abundance, and 2) accounts for non-specific interactions caused by the use of antibodies. Probabilistic modeling of RBP-mRNA association during dynamic conditions was first performed on the RBP HuR during Jurkat T cell activation, and it revealed that the subset of mRNAs associated with HuR changed during the time-course of immune activation **[24]**. These HuR-mRNA subsets when treated as quantitative phenotypes were used to identify drugs that could modulate HuR functionality. This study showed using RIP-chip that investigations of post-transcriptional events can reveal underlying mechanisms of drug action and disease.

In summary the RIP method yields quantitative metrics of RBP-mRNA interactions and uncovers the underlying post-transcriptional mechanisms at play during dynamic biological conditions. However, the major limitation of the RIP method is the inability to directly identify the binding sites of the RBP. Binding site identification can reveal other determinants of binding *in vivo* such as co-occurring motifs for other RBPs, secondary RNA structures important for or antagonistic to binding, and exons that can be alternatively spliced.

1.2 Decoding protein-mRNA interactions: achieving binding site resolution transcriptome-wide

Binding site identification can reveal other determinants of binding *in vivo* such as co-occurring motifs for other RBPs, secondary RNA structures important for or antagonistic to binding, and exons that can be alternatively spliced. The first method developed for identifying the binding sites of RBPs was high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation or HITS-CLIP **[25, 26]**. This method uses 254 nm ultraviolet light to induce covalent crosslinks between RBPs and RNA in cells. Next cell lysates are treated with ribonucleases to digest crosslinked RNAs to small fragments bound by the RBP representing the binding site, and then the crosslinked RBP-RNA complexes are immunoprecipitated with antibodies against the RBP. Polyacrylamide gel electrophoresis is used to stringently purify crosslinked protein-

RNA complexes from non-crosslinked RNA fragments. Crosslinked RBP-RNA fragments are treated with a protease to digest the RBP and then the RNA fragments are used to make cDNA libraries for deep sequencing. A fortuitous consequence of UV crosslinking is it that after treatment with the protease, a small peptide from the bound RBP remains crosslinked to the RNA fragments. This peptide induces point mutations and deletions in the cDNAs due to interference with the reverse transcriptase used to make the cDNAs from the RNA fragment templates **[27, 28]**. These mutations are used to computationally diagnose binding sites of the RBP **[29]**.

Since the development of HITS-CLIP, there has been other CLIP methods developed, each in their own way capitalizing on diagnostic mutations induced by UV crosslinking of RBP to RNA. For example PAR-CLIP (photoactivatable ribonucleoside crosslinking immunoprecipitation) uses metabolic labeling of RNA with 4-thiouridine (4SU), and UV 365 nm to crosslink protein and RNA **[30]**. 4SU that crosslink to protein result in T to C mutations in cDNA made from RNA fragments isolated from the immunoprecipitation of the RBP. These T to C mutations are then used to identify points of contact between RBP and RNA and thus are diagnostic for a binding site. Another technique called iCLIP (individual nucleotide resolution CLIP) takes advantage of the fact that the remaining crosslinked peptide after proteinase treatment can block reverse transcriptase during first strand cDNA synthesis **[31]**. This blockage results in many of

the first strand cDNA being produced without the 5′ adaptor required for PCR amplification. The iCLIP protocol takes advantage of this by replacing one of the intermolecular RNA ligation steps with an intramolecular circularization of the first strand cDNA. The cDNA is linearized to produce the first strand sequence flanked by adaptor sequences for PCR amplification. The point of contact between RBP and RNA is identified from mapped sequenced reads from truncated cDNA by locating the nucleotide immediately upstream (+1) of the mapped read.

1.3 Using CLIP techniques to quantify RBP binding mRNA transcriptome-wide

Efforts to quantify the binding detected in CLIP studies have also included the use of total mRNA-seq as a way to calculate enrichment scores. A study in particular by Kishore and colleagues did HuR HITS-CLIP and PAR-CLIP in which the coverage of HuR binding sites were normalized to coverage of the site in total mRNA-seq. Then they compared the enrichment scores of the binding sites grouped in non-overlapping, rank-ordered sets of 1000 to HuR 7-mer Z-scores from RNAcompete. The RNAcompete Z-score is an enrichment score that measures the preference of 7-mer RNA motifs for binding by an RBP **[32]**. Kishore and colleagues reported good correlation between HuR CLIP binding site enrichment scores and RNAcompete Z-scores, and the number of

crosslinking events **[33]**. However, they did not formally demonstrate a relationship between CLIP enrichment and binding strength.

Another study by Zarnack and colleagues used iCLIP to show transcriptome-wide evidence of direct competition between the RBPs HNRNPC and U2AF65 for binding to U-rich sequences in the introns of pre-mRNA **[34]**. Specifically, they showed that knocking down HNRNPC levels with siRNA resulted in increased binding of U-rich sites by U2AF65 at sites that HNRNPC binds when present. In this study they used changes in crosslinking events at sites in a gene relative to crosslinking events across the gene to quantify protein occupancy. They used this metric to show that there was increased U2AF65 occupancy at certain sites when HNRNPC levels were decreased. Finally they showed how the competition between these proteins protected against the splicing of Alu elements in pre-mRNA. However, it was not formally demonstrated that crosslinking efficiencies were directly related to binding strength.

1.4 Modifications of CLIP methods in order to improve binding site detection

Unlike DNA, mRNAs are present in various copy numbers per cell. Therefore, mRNA abundance must be accounted for in order to quantify binding of the mRNA. Unlike the RIP technique, CLIP techniques at their inception did not incorporate mRNA abundance or background binding into their delineation of binding sites. The initial

focus of CLIP techniques was the identification of RNA binding sites using stringent separation techniques. These steps allowed the separation of the crosslinked RBP-RNA complexes from the non-crosslinked RNA. However, these protocols overlooked the presence of crosslinking background. A study by Friedersdorf and Keene showed that crosslinking background was present when using CLIP procedures. In their study using PAR-CLIP they showed that RNAs were crosslinked to the non-RBP GFP (green fluorescent protein) when it was immunoprecipitated **[35]**. These non-specifically/background crosslinked RNAs were not removed by PAGE (polyacrylamide gel electrophoresis), and thus it was evident that background was present in prior CLIP studies. The authors re-analyzed numerous CLIP experiments that were performed showing that many of the binding sites identified for the RBPs studied overlapped with background sites found in the GFP PAR-CLIP, i.e. there were many false positives. In addition they showed how the background identified by GFP improved RBP binding site identification and motif definition. Following this study by Friedersdorf and Keene, there has been a number of CLIP studies that have performed background crosslinking IPs to improve binding site identification **[36-41]**.

Another obstacle to overcome by CLIP procedures is the inefficiency of UV crosslinking. Early studies had shown variable efficiencies of crosslinking to nucleic acids among the amino acids **[42, 43]**. A study by Sugimoto and colleagues showed

empirical evidence for a uridine bias for RBP binding sites in CLIP studies that used UVC to induce crosslinking **[44]**. In addition, induction of covalent crosslinking of protein and RNA by UV requires that the photoreactive bases and amino acid chains have optimal spatial arrangements relative to each other. This necessity greatly diminishes the ability of UV to crosslink RBPs that recognize RNA secondary structure since these RBPs make contact with the ribose backbone **[45]**.

1.5 mRNPs isolated by RIP do not reassort post-lysis

The use of crosslinking agents were believed to be necessary to ensure that post-lysis reassortment of mRNP complexes did not occur. It was believed that upon lysis of the cell, RBPs were free to dissociate and reassociate with other mRNAs. In a study by Mili & Steitz, it was demonstrated that mRNA bound by an RBP could dissociate from the RBP and reassociate with another RBP upon lysis. The conclusion of Mili & Steitz was that this post-lysis reassortment played a major role in the bound targets observed in RIPs **[46]**. The major difference between the Mili & Steitz experiments and RIP protocols pioneered by the Keene lab is that the former utilized sonication to disrupt RNPs. RIP protocols pioneered by the Keene lab do not utilize sonication because it disrupts protein-RNA interactions and shears nucleic acids and proteins. Instead lysis conditions were intentionally used that would be gentle so as to preserve interactions that were present in the cell post-lysis **[20, 47]**.

A study published by Penalva and colleagues reported the ability to isolate cell type specific mRNPs from co-cultures of mouse endothelial cells and human breast cancer cells. Penalva and colleagues co-cultured endothelial cells expressing the RBP PABP with a G10 epitope tag with breast cancer cells expressing PABP with a FLAG epitope. The cells in the co-cultures were lysed and immunoprecipitations with anti-FLAG antibodies and in parallel with anti-G10 antibodies were performed in order to isolate PABP mRNPs specific to endothelial cells and to the cancer cells. The immunoprecipitations of the PABP mRNP were performed using the buffers and procedures detailed by Tenenbaum and colleagues **[48]**, which does not utilize sonication. The authors showed using RNase protection assays on isolated mRNAs that anti-G10 antibodies enriched PABP bound mRNAs from the mouse endothelial cells, and the anti-FLAG antibodies enriched PABP bound mRNAs from the human breast cancer cell line **[49]**. Thus, post-lysis reassortment does not occur when certain lysis conditions are adhered; for example as detailed in **[47, 48]**.

1.6 RIP and CLIP produce complementary information

Despite the differences between the RIP and CLIP techniques, both yield complementary results. There have been many studies that have utilized both RIP and CLIP to identify functional targets of RBPs **[50-59]**. For example, RIP-chip and PAR-CLIP were used to study the transcriptome-wide targets of the RBP HuR/ELAVL1 **[54]**.

Mukherjee and colleagues showed that mRNAs with higher RIP-chip probability scores for binding by HuR tended to have larger numbers of HuR PAR-CLIP binding sites. In addition, Mukherjee and colleagues showed that both techniques identified targets of HuR that were functional, i.e. HuR targets identified by RIP-chip or PAR-CLIP exhibited reduced expression upon knockdown of HuR levels with siRNA **[54]**. Therefore, RIP and CLIP techniques can identify functional targets of an RBP, and there is a relationship between the number of binding sites for a target by CLIP and RIP-chip probability scores.

1.7 RNA-binding proteins relevant to my studies

The RBPs relevant to my studies all belong to the superfamily of RNA recognition motif (RRM) proteins. RRMs are 80-90 amino acid domains with $\beta\alpha\beta\beta\alpha\beta$ topology, that interact with RNA on the surface of the beta sheet through combinations of electrostatic, hydrogen bonding and base-stacking interactions **[60]**. RRMs are comprised of two conserved consensus sequences called RNP1 and RNP2, and are sufficient for binding RNA with wide ranges of affinities and specificities **[61-63]**. A single RRM can recognize anywhere from 4 to 8 nucleotides dependent on the use of exposed loops and other secondary structures absent from the canonical structure **[60, 64]**. Also proteins with multiple RRM domains recognize longer stretches of RNA sequences often times with higher affinities **[60]**.

In addition some RRMs interact with proteins suggesting a role for protein-protein interactions in the binding of RNA. There have been many examples of inter-domain interactions between two RRMs. These interactions can be independent of RNA binding, for example two RRMs in hnRNPA1 **[65, 66]**. They can also be RNA dependent, for example the cases of nucleolin, sex-lethal, PABP, and HuD **[67-70]**. Interestingly, there can be interactions between two RRM domains resulting in the cooperative binding of RNA. For example the RRM domains of two U1A proteins cooperate to bind with higher affinity to the polyadenylation inhibition element (PIE) RNA **[71]**. The affinity of this complex was predicted to be orders of magnitude greater than if binding was additive. Also, the cooperative binding of the PIE RNA was necessary for inhibition of polyadenylation suggesting that these types of interactions can be important for regulation of mRNA **[72]**.

In summary, RRMs are RNA binding domains whose modular design bestows the ability to bind RNA with high affinity and specificity. RRM proteins in general bind to single-stranded regions of RNA and there are some examples of inter- and intramolecular interactions between RRMs that are important for functional outcomes. The three RBPs relative to my study; ELAVL1/HuR, RBM38, and TRA2 all contain RRMs, and are described below.

1.7.1 ELAVL/Hu proteins

HuR is one of four members of the ELAV family of RBPs that also includes HuB, HuC, and HuD. HuR is ubiquitously expressed in the nucleus of human cells while the other family members are neuron-specific and predominantly cytoplasmic. HuR and its family members have three RRMs, with the second and third RRMs separated by a flexible linker region.

The neuronal family member HuB was the first to have its binding preference elucidated by Levine and colleagues **[73]**. They distinguished through *in vitro* selection procedures that HuB predominantly bound to RNA containing stretches of 3 to 5 uridylate residues separated by other nucleotides. It was also shown in this study that HuB could bind the the 3'UTRs of c-fos, c-myc, and GM-CSF that contain uridylate rich sequences. Deletion of a 29 nucleotide, U-rich sequence from a 181 nucleotide sequence in the *c-myc* 3'UTR reduced binding of HuB to undetectable levels. This provided evidence that the U-rich sequences were necessary for binding by HuB. It was later shown by Myer and colleagues through UV crosslinking and electrophoretic mobility shift assays that HuR binds with high affinity to AU-rich sequences **[74]**. HuR RIP-chip and bioinformatic analysis were performed by de Silanes and colleagues revealing that most of the transcriptome-wide targets of HuR contained U-rich sequences that were predominantly in 3'UTRs **[75]**. Meisner and colleagues further refined HuR's binding

preference using computational and biochemical assays by deducing a NNUUNNUUU consensus motif, and that this motif must be present in single stranded conformation **[76]**. HuR's U-rich consensus motif and the requirement of its single-stranded presentation have been corroborated in later studies **[32, 77, 78]**.

The discovery of HuB's recognition of A/GU-rich sequences in short lived mRNAs encoding growth factors lead to the hypothesis that HuB's binding would have an effect on mRNA stability **[73, 79]**. Indeed, an investigation by Jain and colleagues revealed that HuB could bind to the U-rich sequences in 3'UTR of GLUT1 mRNA resulting in both increased stability and translation of the mRNA **[80, 81]**. It was later shown that overexpression of HuR in mouse L929 cells increased the stability of beta-globin reporters containing AREs **[82]**, stabilize c-fos and VEGF mRNAs **[83, 84]**, and can increase the translation of p53 mRNA in response to ultraviolet light irradiation **[85]**.

A mechanism for HuR's function on mRNA's is at least in part due to its antagonism of microRNA (miR) function. One of the first examples of this was the reversal of miR-122 repression of CAT-1 mRNA by stress-induced HuR **[86]**. The interpretation was that AGO-miR-122 sequestered CAT-1 mRNA into processing bodies, but when cells are stressed the mRNA is bound by HuR which facilitates binding by polysomes and translation of the mRNA. Another mechanism for HuR's antagonism of miR function was that HuR physically blocks the binding of miRs at nearby sites.

Mukherjee and colleagues showed by integrating HuR PAR-CLIP and transcriptomics that mRNAs possessing HuR binding sites within 10 nucleotides of a miR site were in general more unstable upon knockdown of HuR levels than mRNAs having HuR and miR sites separated by greater distances **[54]**. The antagonism of miR function by HuR binding to nearby sites was also observed in when macrophages were activated **[87]**. On the contrary, HuR's binding of mRNAs can also have repressive effects. Evidence supporting this was reported by Kim and colleagues when it was revealed that HuR could recruit miR-let-7 to MYC mRNA 3'UTR resulting destabilization and translation repression of the mRNA **[88]**.

1.7.2 RBM38

RBM38/RNPC1 is a sequence specific RNA-binding protein that contains one RRM, and no other defined domains. It is expressed in many species of multicellular eukaryotes, from nematodes up to humans. Its first characterization was in a screen to detect genes that were regulated by tumor suppressor p53 **[89]**. Since its identification as a target protein of p53, there has been further reports on RBM38's role in the post-transcriptional regulation and translation of mRNA, the post-translational modifications that modulate its function, differentiation, development, and cancer/tumor suppression. In mammals, RBM38's expression is greatest in differentiating erythocytes **[90]**, and in differentiated muscle cells **[91]** where it is understood to be involved in tissue-specific

alternative splicing. RBM38's function appears to be most important to the proper development of blood cells from the hematopoietic lineage. For example mice deficient in RBM38 exhibited increased p53 expression, abnormal hematopoiesis, accelerated aging, and spontaneous tumor (mostly lymphomas) formation **[92]**.

RBM38 binds with strongest affinity to $(GU)_n$ motifs in RNA **[90, 93]**, although other studies have reported binding to AU-rich elements, and other U-rich sequences **[94, 95]**. RBM38 reportedly binds its targets at various locations in mRNA such as the 5'UTR, 3'UTR, and introns. RBM38's binding to the UTRs have been implicated in modulating its targets stability or translation efficiency **[94, 96-103]**, while binding of introns can cause alternative splicing **[90]**. Interestingly, a number of studies have shown cooperative binding of mRNA by RBM38 and other RBPs such as HuR/ELAVL1 in mammalian cells, and RBFOX in *C. elegans*. Specifically, one study reports RBM38 and HuR cooperatively binding p21/CDKN1A mRNA in the 3'UTR thus increasing mRNA stability **[94]**. Another study in *C. elegans* reported RBM38 (Sup-12) and RBFOX (Asd-1) binding in the introns of egl-15 mRNA causing alternative splicing of specific exons necessary for protein expression **[104]**. The studies in *C. elegans* were extended by NMR solution structures of Sup-12-Asd-1-RNA complex showing the physical interactions necessary for the formation of the complex **[105]**. However, since these studies were

limited to testing a small number of selected mRNAs, they do not establish cooperative binding of mRNA by RBM38 with other RBPs as a general mechanism.

1.7.3 TRA2

Tra2 proteins are highly conserved in the animal kingdom; they are found in vertebrates and invertebrates. The first TRA2 protein to be studied was the one expressed by *Drosophila*. In *Drosophila*, TRA2 is involved in female sex determination, and roles in spermatogenesis have been reported **[106-108]**. Humans encode two TRA2 homologs; TRA2A **[109]**, and TRA2B **[110]**. The TRA2 proteins have one RRM flanked by serine-arginine (SR) rich domains, and are located predominantly in the nucleus. TRA2B is implicated in many human diseases including cancer **[111]**, spinal muscular atrophy **[112]**, Alzheimer's disease **[113]**, and fronto-temporal dementia and parkinsonism linked to chromosome 17 **[114]**.

The TRA2 proteins are not required for the constitutive splicing of exons, but activate splicing in a sequence-dependent manner **[115]**. Tacke and colleagues also identified the sequence of preference for both TRA2 proteins to be GAA repeats using SELEX **[115]**. The transcriptome-wide binding of TRA2 proteins have also been studied using RIP-seq, HITS-CLIP, and iCLIP and all studies detect transcripts with AGAA sequences being most abundant in the binding sites **[116-118]**. HITS-CLIP and iCLIP studies showed that TRA2 proteins bound predominantly to coding sequence exons

[116, 118]. However, it was reported that the TRA2B protein could bind to CAA sequences using a different mode of binding to RNA **[119]**.

1.8 The focus of my dissertation research

Although the field has developed methods to identify the targets of RBPs at the whole transcript and binding site levels, there are still questions about binding and regulation that remain. For example what are the differences between weak and strong binding events? How do we distinguish weak versus strong binding events? Can we distinguish other determinants of binding in addition to RNA sequence motifs? The RIP method quantitatively identifies the targets of RBPs, but does not directly identify binding sites. CLIP methods have been improved to identify binding sites, and attempts have been made to gain quantitative binding information from CLIP experiments. However, these experiments have not definitively shown that differences in binding are due to differences in binding strength. In addition, the fact that mRNAs are expressed in multiple copies per cell means that there is heterogeneity to the binding that is observed. For example, when binding sites in mRNA are identified, are all binding sites bound in all the copies, or is it a subset of the sites? In theory, we can obtain a better idea of which sites are likely to be most bound on a copy number basis by including transcript abundance in our calculations of binding sites. Ultimately, this should allow for better

relative comparisons of "pseudo-affinities" for RNA binding sites of RBPs in the cell, and help in identifying binding sites that are modulated in cells in response to stimuli.

Despite their different methodologies, RIP and CLIP can produce complementary data when used to study the same RBPs under similar conditions. Therefore, the focus of my dissertation research has been the development of a technique that will identify the binding sites of RBPs and provide quantitative metrics of binding that approximate binding strength of the RBP. I used the RBP HuR/ELAVL1 as a test case for developing the principles of this technique that I call digestion optimized RNA immunoprecipitation with deep sequencing or DO-RIP-seq. I found that DO-RIP-seq binding site quantitation can approximate the relative binding strength of HuR at RNA sites. In addition, DO-RIP-seq can also assign scores for whole transcript targets similarly to scores yielded by RIP-chip/seq studies. The effect of binding to sites in the transcript is observed on the whole transcript. So being able to relate binding sites to targeting of the whole transcript will allow one to better connect binding to functional outcomes at the whole transcript level. I also used DO-RIP-seq to study the transcriptome-wide binding of RBM38. I have found that RBM38 binding sites had a significant co-occurrence with the sequence motifs of HuR and TRA2. Sites with the co-occurring RBM38 and TRA2 or RBM38 and HuR sequence motifs had higher binding scores in general indicating the potential for stronger binding or protein-protein

interactions at these sites. Also, sites where RBM38 and HuR binding overlapped were more conserved than non-overlapping RBM38 sites.

2. Determination of optimal RNA digestion conditions with Micrococcal nuclease

2.1 Introduction

My goal was to develop a technique that could identify the binding sites without crosslinking protein and RNA, and yield quantitative metrics of binding to the RNA site. In order to do this I melded technical aspects of RIP and CLIP protocols. For example the cell lysis buffer I selected is from the RIP protocol that was designed for gentle lysis of cells to preserve RNP complexes, and minimize RNA degradation by RNases **[47, 48]**. I borrowed from CLIP procedures the use of RNases to digest protein bound RNA into fragments representing the protein's binding site on the RNA **[30, 31, 41, 120]**. First, the optimization of digestion conditions using a ribonuclease or nuclease was required. The nuclease of choice was Micrococcal nuclease (MNase) because it requires Ca^{2+} for activity, and is inactivated when Ca^{2+} is chelated **[121]**. This feature of MNase makes it ideal for the optimization of RNA digestion, and also controlling the extent of digestion. Controlling the extent of nuclease digestions is important in order to limit the sequence biases of RNA fragments produced which become more pronounced when the digestion is extensive **[33]**. In addition, MNase has optimal activity at 37 °C but is active at temperatures as low as 2 °C and so can be used at sub-optimal temperatures **[122]**. I selected the RBP HuR as a test case because of our labs expertise in studying this protein

and its well characterized binding specificity. Given that HuR binds to U-rich RNA sequences in mRNA and some long non-coding RNA, I used it as a test case for identifying conditions that would yield RNA fragments that predominantly mapped to mRNA, with the occasional U-rich motif.

There were also size constraints to be considered for the RNA fragments produced. The RNA fragments isolated from immunoprecipitations of the RBP and with non-specific antibodies would be used to make cDNA libraries for deep sequencing. After sequencing, these cDNAs would be mapped to the human genome. The sequences must map uniquely to the genome in order to unambiguously identify the sites bound by the protein. However not all sequences will map uniquely to the genome because of their size. The mappability of a sequence is improved by increasing the read length (i.e. RNA fragment size) **[123]** and thus places a lower constraint on the RNA fragment sizes used for mapping the binding sites. In order to set an upper size limit for RNA fragments utilized for mapping binding sites, I reasoned that this limit would be determined by cDNA library and deep sequencing protocols used. I selected the small RNA library making protocol first developed by Hafner and colleagues to make cDNA libraries from ~21 nucleotide small RNAs called microRNAs **[124]**. The small RNA library protocol of Hafner and colleagues has been adopted, modified, and made commercially available by multiple sources.

In order to sequence the cDNA libraries from optimization experiments, I used a sub-cloning strategy where cDNA libraries are ligated into a DNA vectors, and then used to transform *E. coli*. Transformed *E. coli* will express antibiotic resistance gene encoded by the DNA vector containing the inserted cDNA, and so their growth antibiotic growth plates will select colonies that have sub-cloned the DNA vector with cDNA sequences. These colonies can be picked, and the cDNA extracted sequenced using Sanger methods. This strategy was first utilized by Ule and colleagues to identify binding sites of the RBP NOVA using CLIP **[25]**.

2.2 Materials and Methods

Two strategies were employed to determine conditions that would yield RNA fragments representing the binding sites of RBPs: 1) determining conditions from total RNA fragments isolated from nuclease treated lysates, and 2) determining conditions from RNA fragments extracted from immunoprecipitations of RBPs from nuclease treated lysates. Below I describe both strategies.

2.2.1 Cell culture and lysis

Human embryonic kidney 293 (HEK293) cell line was obtained from the Duke Cell Culture Facility which maintains clones first obtained from the American Type Culture Collection (ATCC). HEK293 cells were grown in humidified incubators at 37 °C, 5% CO_2 in 15-cm dishes until they were 80-90% confluent (approximately 12 x 10^6 cells

per dish). The growth media used was Dulbecco's Modified Eagle Medium (Thermo Fisher, cat no. 11965092) supplemented with 10% (v/v) fetal bovine serum. Cells were harvested by first removing the growth media, and then harvested directly into cold phosphate buffered saline, pH 7.4 (PBS) by scraping. Cells were washed once with ice cold PBS, and then lysed on ice by resuspending and repeated pipetting with 1.5 cell pellet volumes of polysome lysis buffer (10 mM HEPES pH 7.0, 100 mM KCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$, 1X cOmplete™ protease inhibitor cocktail (Roche), 100 Units/ml RNaseOUT (Life Technologies), 0.5% IGEPAL CA-630, and 1 mM DTT). Lysates were incubated on ice for 5 min and then stored at -80 °C.

2.2.2 Preparation of magnetic protein G beads

1.5 mg of magnetic protein G beads (Life Technologies) were prepared by removing the bead storage buffer, and then resuspending the beads in 200 µl of NT2 buffer (50 mM Tris-HCl, 1 mM MgCl2, 150 mM NaCl, 0.05% IGEPAL). The protein G beads were incubated with saturating amounts of mouse anti-HuR (3A2) monoclonal antibodies from hybridoma supernatant, or with 1 µl (50 µg) of normal mouse serum (pre-immune serum, Jackson ImmunoResearch Laboratories) antibodies at room temperature for 20 minutes or overnight (~18 h) at 4 °C. Before antibody bead complexes were used for immunoprecipitation, they were washed three times with 500 µl of cold NT2 buffer.

2.2.3 Optimization of Micrococcal nuclease conditions for partial digestion of RNA by analyzing total RNA

Three 15-cm dishes of HEK293 lysates were prepared using conditions described above. Frozen cell lysates were thawed on ice and cleared by centrifugation at 15,000 g, 4 °C for 15 min. 10% of the pooled lysate (undigested input) volume was set aside to measure total RNA concentration before treatment with Micrococcal nuclease (New England BioLabs). Total RNA was extracted from the undigested input using the TRIsure reagent (Bioline) and the manufacturer's protocol. The total lysate was split into seven equal volumes. The amounts of MNase and the durations of the reaction were tested at a constant temperature of 30 °C. The conditions tested in the format MNase (gel units/ μg total RNA) - duration (min.) - temperature (°C) were 0-0-0, 0-30-30, 100-10-30, 200-10-30, 100-30-30, 200-30-30, and 200-120-30. Reactions were stopped by adding EGTA at a molar ratio of 2:1 with respect to $CaCl_2$. Total RNA was extracted using TRIsure reagent per the manufacturer's protocol and submitted to the Duke center for genomic and computational biology for analysis on Agilent Bioanalyzer.

2.2.4 Optimization of Micrococcal nuclease conditions for partial digestion of RNA by analyzing immunoprecipitated products

Six 15-cm dishes of HEK293 lysates were prepared using conditions described above. Frozen cell lysates were thawed on ice and cleared by centrifugation at 15,000 g, 4 °C for 15 min. After clearance the cell lysates were pooled together. 10% of the pooled

lysate (undigested input) volume was set aside to measure total RNA concentration before treatment with Micrococcal nuclease (New England BioLabs). The total RNA was extracted from the undigested input using TRIsure reagent and protocol (Bioline), and the total RNA measured using a NanoDrop spectrophotometer. The remaining lysate was divided into six equal volumes, and each was increased to a final volume of 300 μl using polysome lysis buffer. Next the lysates were incubated at 30 °C for 5 minutes with increasing amounts of Micrococcal nuclease and also without Micrococcal nuclease. Ratios of Micrococcal nuclease (gel units) to total RNA (μg) tested were 0, 1, 5, 18, 73, and 294 gel units/μg. The Micrococcal nuclease was inactivated by adding EGTA to a final concentration of 10 mM (a 2:1 molar ratio of EGTA to $CaCl_2$).

2.2.5 Immunoprecipitations of HuR to isolate bound RNA fragments

Ten 15-cm dishes of HEK293 lysates were prepared using conditions described above. Frozen cell lysates were thawed on ice and cleared by centrifugation at 15,000 g, 4 °C for 15 min. After clearance the cell lysates were pooled together. 10% of the pooled lysate (undigested input) volume was set aside to measure total RNA concentration before treatment with Micrococcal nuclease (New England BioLabs). The total RNA was extracted from the undigested input using TRIsure reagent and protocol (Bioline), and the total RNA measured using a NanoDrop spectrophotometer. Next the remaining pooled lysate were incubated at 30 °C for 30 minutes with either 200 gel units of

MNase/μg total RNA (selected from the Bioanalyzer analysis), or for 5 min with 30 gel units/μg (selected from the analysis of immunoprecipitated RNA fragments). The micrococcal nuclease was inactivated by adding EGTA to a final concentration of 10 mM (a 2:1 molar ratio of EGTA to $CaCl_2$).

Immunoprecipitations were performed with magnetic protein G beads coupled to mouse anti-HuR (3A2) monoclonal antibodies. Immunoprecipitations were performed in a final reaction volume that was 8X the lysate volume using NT2 buffer supplemented with 100 units/ml RNaseOUT, 20 mM EDTA, and 1 mM DTT. Immunoprecipitation reactions were incubated at 4 °C for 2 h with constant mixing. Afterwards the bead complexes were washed 4 times with ice cold NT2 buffer, and the RNA was extracted using the TRIsure reagent per the manufacturer's protocol.

Dephosphorylation, radiolabeling and electrophoresis of RNA fragments.

Digestion of RNA with MNase produces fragments with 3′-PO_4^{3-} and 5′-OH. In order to ligate the sequencing adapters, 3′ ends have to be dephosphorylated and 5′ ends radiolabeled with P^{32}. Extracted RNA was treated with 1 unit of recombinant shrimp alkaline phosphatase (rSAP, New England BioLabs) in 10 μl reaction volumes containing 1X CutSmart® buffer (New England BioLabs). The rSAP was inactivated by incubating at 65 °C for 5 min, and then the reactions cooled on ice. Next, the RNA was radiolabeled with 0.5 μl of 6000 Ci/mmol [γ-P32]ATP (Perkin Elmer), 1X T4

polynucleotide kinase buffer (New England BioLabs) and 10 units of T4 polynucleotide kinase (New England BioLabs) in a final reaction volume of 20 μl for 30 min at 37 °C. One volume of 2X gel loading buffer II was added to each reaction, and then they were incubated for 30 s at 95 °C and then cooled on ice. The reactions were loaded onto 15% Criterion™ TBE-Urea polyacrylamide gels (Bio-Rad), and electrophoresis performed. Gels were exposed to phosphor imager screens and then the screens scanned on a phosphor imager.

Extraction of RNA fragments from polyacrylamide gel. An "actual size" image of the gel was used as an aid to cut RNA fragments corresponding to 25-70 nucleotides in length cut from the gel, and extracted overnight by constantly mixing the gel slices in 400 μl 0.4 M NaCl at 4 °C. The supernatants containing the radiolabeled RNA fragments were extracted and transferred to new microcentrifuge tubes and washed by adding 200 μl 0.4 M NaCl, and constantly mixing at 4 °C. The supernatant from the washes were pooled with the others, 30 μg of GlycoBlue™ added to each followed by 1 ml of 100% ethanol to each, and then the RNA fragments precipitated by incubating at -20 °C for 1h or -80 °C for 30 min. The RNA fragments were centrifuged at 17,000 g, 4 °C for 20 min, and then the supernatants removed. RNA pellets were washed once with 80% ethanol, centrifuged at 10,000 g, 4 °C for 5 min, the supernatants removed and then the pellets air-dried.

Preparation of cDNA libraries from size-selected RNA fragments. My preferred method for ligating sequencing adapters, making the first strand cDNA, and amplifying the final cDNA library was by using the NEBNext® small RNA library prep protocol (New England BioLabs). This protocol allows the sequential ligation of a DNA oligonucleotide to the 3′ ends of the RNA fragments followed by the liagtion of a RNA oligonucleotide to the 5′ ends, catalyzed by T4 RNA Ligase 2 (truncated, K227Q mutant, New England BioLabs) and T4 RNA Ligase 1 (New England BioLabs) respectively. A custom 3′ adapter containing a pool sequences with random 5 nucleotide sequences at the 5′ end was used instead of the one provided. The random nucleotides in this adapter serves as unique molecular identifier (UMI) of biologically duplicated sequences versus PCR produced duplicates. First strand cDNA is made in a reaction catalyzed by Moloney Murine leukemia virus reverse transcriptase, and using a DNA primer to allow first strand cDNA to be polymerized.

In order to determine the minimum number of PCR cycles required for production of the final cDNA library, one-tenth of the first strand cDNA was used in PCR using the kit components provided. Small aliquots of the PCR were taken at different cycle numbers, and analyzed by agarose gel electrophoresis. Cycle numbers at which a faint PCR product was visible was used to make the final cDNA libraries. After completion of the PCR, the cDNA products were extracted by phase separation with

phenol-chloroform-isoamyl alcohol, and precipitated with ethanol. The PCR yields amplified cDNA fragments from the immunoprecipitated RNA fragments and cDNA adapter concatemers resulting from the ligation of 3′ adapter to the 5′ adapter. This required that purified cDNA be separated by agarose gel electrophoresis using 4% low melt agarose gels, and cDNA fragments of the desired size be cut and extracted from the gel using the QIAquick gel extraction protocol (Qiagen).

Validation of cDNA library sequences. A small volume of the cleaned up cDNA libraries were ligated into pGEM-T DNA vectors and used to transform *E. coli* (DH5alpha strain). The *E. coli* were plated on LB agar plates containing ampicillin, and incubated overnight at 37 °C. Transformants were picked, and PCR using primers to the T7 and SP6 promoters performed to amplify cDNA sequences inserted into the pGEM-T vector. Amplified sequences were cleaned up using QIAquick PCR purification protocol (Qiagen) and submitted to Eton Biosciences for Sanger sequencing. Sequences were mapped to the human genome using the BLAT tool in the UCSC genome browser.

2.3 Results

The most prominent RNA species observed on the Bioanalyzer were 28S and 18S rRNAs. Messenger RNAs are much less abundant than rRNA and their digestion profiles were not readily distinguished on the Bioanalyzer **(Fig. 1A)**. The RNA fragments from the digestion optimizations that included IPs of HuR displayed the

eventual enrichment of fragments between 20 and 70 nucleotides in length **(Fig. 1B)**. Given these results I selected two conditions (one from each optimization strategy) as a test case of a DO-RIP-seq run where the sequences from the IPs would be analyzed. I expected that proper digestion conditions would yield sequences mapping predominantly to mRNA, and that some would contain sequences resembling the known motif recognized by HuR. When 200 gel units of MNase per μg of total RNA was used, the cDNA sequenced from *E. coli* transformed with libraries made from the HuR IP RNA fragments all mapped predominantly to rRNA genes. None of the sequences mapped to mRNAs **(Table 2)**. However, sequences from cDNA libraries prepared with RNA fragments from HuR IPs from lysate treated with 30 gel units per μg for 5 min **(Fig. 2)** all mapped to mRNA-coding genes **(Table 1)**. In addition many of these cDNA sequences contained HuR-like motifs.

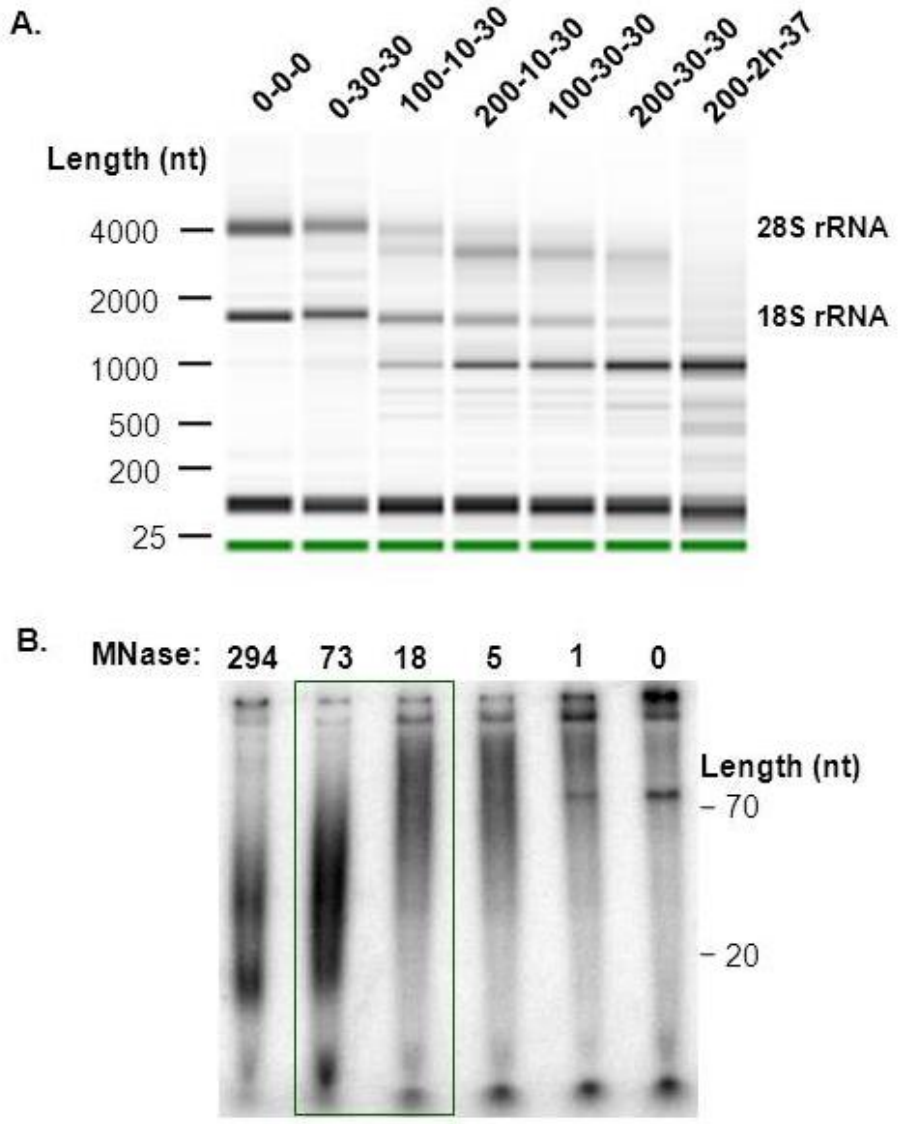

Figure 1: A. The gel image from the Agilent Bioanalyzer displaying the lengths of RNA fragments (nucleotides, nt) produced from the treatment of HEK293 cell lysates with various amounts of Micrococcal nuclease (MNase, gel units to total RNA (μg)), and for various durations and temperatures. B. The radiolabeled RNA fragments extracted from immunoprecipitations of HuR from cell lysates treated with various amounts of MNase were separated on a 15% TBE-Urea polyacrylamide gel. The green box highlights the range for optimal digestion conditions.

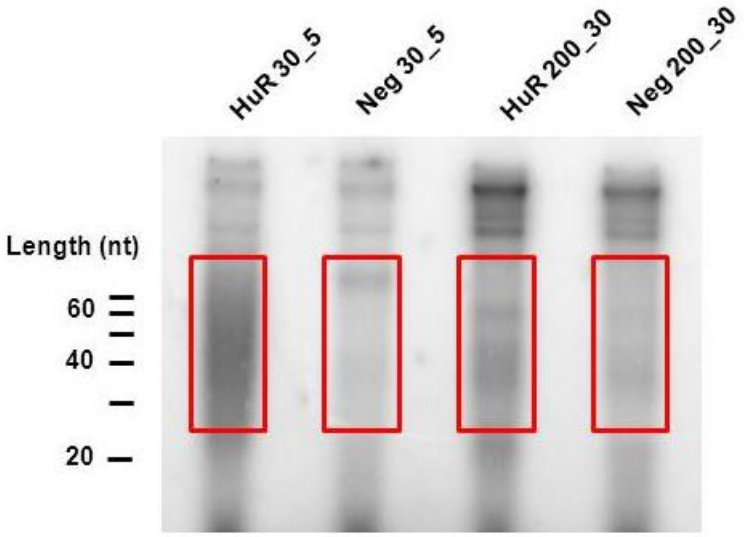

Figure 2: RNA fragments from HuR and negative (Neg) control DO-RIP that were separated by PAGE. Lysates were treated with 200 gel units MNase per μg total RNA for 30 min (HuR/Neg 200_5) or 30 gel units per μg for 5 min (HuR/Neg 30_5). Red rectangles demarcate the approximate region of the gel cut for extracting the RNA. [nt = nucleotides].

Table 1: HuR DO-RIP sequences from optimal digestion conditions

>CCT8_CDS_chr21:30,434,842-30,435,033_-
TTAGTGAGGCTAAACTCAAAATGGGATCTCCGAAGACTTTGTAA
>HSPA5_CDS_chr9:128,000,571-128,000,591_-
ACAGGTGACCTGGTACTGCTT
>PAIP1_CDS_chr5:43,535,713-43,535,744_-
AAAGTTGACAGGATCAGTTTTGGAAGATGCTTGG
>SON_CDS_chr21:34,929,588-34,929,612_+
AGTCAGGAGGAGCTACTATAGAAGA
>LPGAT1_3UTR_chr1:211919329-211919357_-
TTGTTTTCTTTGTTCTTTTTTTGAAGTCA
>RAN_3UTR_chr12:131,360,755-131,360,795_+
ATCTTGTTTGTTACTGTCATTCCCATTCCTTTTCGTTTAGA
>SS18_3UTR_chr18:23597210-23597238_-
TATGTTTTTCCCTTTTTTATGTTGGTTGA
>SUB1_3UTR_chr5:32,602,122-32,602,167_+
ATTAAGCTTTACTGTTATAGTAGGTAATATGGTTAGTTTGTAGGGA
>DNAJB11_intron_chr3:186,296,085-186,296,130_+
TTTTTGTTTCTTCCAGTTTTATTCAATTCCAAACAGCGCTTTTATG
>NCOA7_intron_chr6:126,209,995-126,210,024_+
AGTTTCCGTTTTGTTTTTTGCCTTTCCTCC
>chrM:14,678-14,726
ACAACGATGGTTTTTCATATCATTGGTCGTGGTTGTAGTCCGTGCGAGA
>PPP6R3_5UTR_chr11:68,287,053-68,287,076_+
ACCTGTAATGGGCAAGGAGCCACA
>DNAJC14_anti-sense_chr12:56,216,465-56,216,505_+
ACACTCAGCACAGTATCTGGCACTCTTAGGTTCCCGGTCCA

Table 2: HuR DO-RIP sequences from over-digested conditions

>rRNA_chrUn_gl000220:115,785-115,844_-
GCAACGGAGGCCATCGCCCGTCCCTTCGGAACGGCGCTCGCCCATCTCTCAGGACCGACT
>rRNA_chr17:33,478,208-33,478,234_-
ATATTAGTCAGCGGAGGAAAAGAAACT
>rRNA_chr21:9,827,183-9,827,220_-
GCACGCATCCCCCCCGCGAAGGGGGTCAGCGCCCGTCG
>rRNA_chrUn_gl000220:110,738-110,773_-
GTACAAAGGGCAGGGACTTAATCAACGCAAGCTTAT
>rRNA_chr12:20,704,449-20,704,484_-
AATTCCGATAACGAACGAGACTCTGGCATGCTAACT
>rRNA_chrUn_gl000220:109,717-109,755_-
AATAGCGTATATTAAAGTTGCTGCAGTTAAAAAGCTCGT
>rRNA_chr17:33,478,235-33,478,275_-
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATTTAAGC
>7SK_chr6:52,860,434-52,860,469_-
CCCTGGCGATCAATGGGGTGACAGATGTCGCAGCCA
>rRNA_chrUn_gl000220:110,523-110,559_-
GCTCAATCTCGGGTGGCTGAACGCCACTTGTCCCTCT

2.4 Discussion

The optimization of RNA digestion conditions for DO-RIP-seq is crucial to enrich the binding sites of the RBP. I learned from the digestion optimization experiments that it is more informative to analyze the digestion profile of the IP RNA fragments than analyzing the digestion profile of total RNA from digested lysates. It was also evident that the over digestion of the RNA by MNase resulted in the depletion of mRNA fragments, and the increase of non-specifically bound, abundant non-coding RNAs like rRNA. This was likely the case because HuR is reported to bind mRNA and some long non-coding RNA but not rRNA. Therefore some prior knowledge of the RBP's targets is valuable for optimizing RNA digestion conditions to yield binding sites. Though cDNA sequenced from transformants represent a miniscule proportion of all sequences in a

cDNA library, it may serve as a useful validation of the identity of sequences obtained for mapping the RBP's binding site. In conclusion, I developed a protocol that could isolate partially digested RNA bound by HuR, and synthesize a cDNA library from these RNA fragments for deep sequencing.

3. Mapping and quantitatively defining the binding sites of HuR transcriptome-wide with DO-RIP-seq

3.1 Introduction

With digestion conditions worked out for obtaining HuR bound RNA fragments, I moved forward with the performance of three biological replicates of HuR DO-RIP-seq in HEK293 cells each with negative controls. The goals of this experiment were to obtain cDNA libraries that would be deep sequenced and then used to quantify the binding of HuR across the transcriptome, and obtain binding sites exhibiting known features of HuR binding sites. The binding sites of HuR would be determined using a local enrichment approach by normalizing the local abundance of sequenced reads from HuR IPs to local abundances exhibited by the negative control IP. The question is can quantifying binding sites approximate the relative binding strength at sites? Also important is obtaining improved transcriptomic coverage from DO-RIP-seq. Improving the transcriptomic coverage will allow for a greater ability to quantify all binding sites in expressed transcripts. Accomplishing this would allow bound sites and non-bound sites to be distinguished. The non-bound sites being referred are those containing the sequence motif recognized by HuR yet is not bound in the biological conditions being studied. By distinguishing bound and non-bound sites, the determinants of HuR binding, and by extension other RBPs can be identified. The work described here was

carried out in partnership with Matthew Friedersdorf and has been published in the journal *RNA* **[125]**.

3.2 Material & Methods

3.2.1 HuR DO-RIP-seq

Ten 15-cm dishes of HEK293 cells were grown in conditions described in section **2.2.1** for each of the three DO-RIP-seq experiments. Cells were harvested from different passage numbers and several weeks apart. Magnetic protein G beads were coupled to mouse anti-HuR antibodies and in parallel to normal mouse serum antibodies as described in section **2.2.2**. Cell lysates were thawed, and cleared as described in section **2.2.4**. The total RNA concentration was determined from 10% of the pooled lysate and then the remainder treated with 30 gel units MNase per μg of total RNA for 5 min at 30 °C. MNase was inactivated with EGTA as described in section **2.2.4**. HuR DO-RIP-seq and negative DO-RIP-seq were performed in parallel and then the RNA fragments extracted. Extracted RNA fragments were dephosphorylated, radiolabeled and then size separated by PAGE as out lined in section **2.2.5**. RNA fragments corresponding to oligonucleotide lengths of 25-70 were cut and extracted from the gel, and then used to make cDNA libraries using the NEBNext small RNA library preparation protocol as described in section **2.2.5**. The cDNA sequences were validated and submitted to the Duke Genomic and Computational Biology core for 100 bp single read sequencing on Illumina Hi-seq 2500.

3.2.2 Processing of raw sequenced reads and data analysis

As mentioned section **2.2.5**, custom adapter sequences that contained five random nucleotides were used to distinguish PCR artifact from biological duplicate sequences. To process raw reads from the DO-RIP-seq cDNA libraries, adapter sequences up to, but not including, the five-nucleotide UMIs were removed using Cutadapt. These trimmed reads still containing the UMIs were then collapsed to unique sequences. The UMIs were removed and the reads were mapped to the human genome (hg19) using TopHat2 **[126]**. Whole transcript association with HuR was empirically determined using a "RIP-Seq-Like" or RSL score. This was achieved by first calculating reads per million mapped (RPM) for each known gene (Gencode v19, excluding intronic reads) in the HuR and negative libraries. Then the difference between the logarithmic means for the RPMs of each gene in HuR and the negative were defined as the RSL score of the gene.

Binding sites of HuR were identified by binning uniquely mapped HuR and negative reads were in 5 nucleotide (nt) intervals across the genome. In the cases where either there were reads present in the HuR interval or the negative interval but not in the other, a pseudo-count of one was added to the interval with zero reads according to Laplace's rule of succession. Reads were normalized for library size (per million reads) and $\log_2$ transformed. Reproducible enrichments of reads in HuR intervals versus matched sample negative intervals were determined by single-tailed paired t-test. A

high pass filter was applied where an interval must have at least one sample (HuR or negative, including all replicates) with greater than six reads. The ratios of HuR reads to negative reads for each 5 nt interval (difference of log transformed normalized counts) were calculated for the mean enrichment across the replicates. Histograms of this log ratio produced a characteristic bimodal pattern of enrichment, with the two apparent distributions consisting of intervals enriched or depleted in HuR versus negative **(Fig. 5A)**. To determine the probabilities of each interval belonging to the HuR enriched distribution, mixture models were generated using the 'mixtools' package in R as previously described **[24]**. The posterior probabilities generated from the mixture model were used to calculate a log of odds ratio or LOD score (log10 of ratio of posterior probabilities). HuR binding sites were defined as three or more consecutive 5 nt intervals with LOD > 0 and $p < 0.05$.

Transcriptome coverage comparisons. To calculate the transcriptome coverage first, all transcripts were defined as the known gene present in Gencode version 19. These transcripts were filtered for expression in HEK293 cells using mRNA-seq data **[127]**. Expression was categorized as 'detectable' if the transcript had an FPKM>0, 'expressed' if the FPKM>0.1 and 'abundant' if the FPKM>1. Percent coverage of each expression category was then calculated for DO-RIP-seq and PAR-CLIP data sets (HuR and negative IPs from both) using bedtools **[128]**.

Identifying enriched gene sets and binding site relationships. Functional enrichments were calculated using GSEA **[129]**. A gene set's enrichment was considered significant if it had FDR q-value < 0.05 or family wise error rate (FWER) < 0.1. Overlapping sites were defined as predicted 7-mer miRNA sites (predicted by TargetScanHuman 5.1 **[130-132]** within 10 nucleotides of a HuR binding site. The frequency of overlap was defined as ratio of miRNA seed matches overlapping HuR sites relative to the total number of miRNA seed matches in the genome. Shuffled miRNA sites were produced using only miR sequences shuffled on 3'UTR sequences from protein coding genes.

Binding site saturation analysis. As described previously **[35]**, the fraction of binding sites observed was calculated using random samples of mapped reads at depths equivalent to 10%, 30%, 50%, 70%, and 90% of each of the HuR replicate libraries. Binding sites were then calculated as described above for each sampling depth. For each depth these newly determined binding sites from the sampled reads were compared to the binding sites determined from the whole libraries to calculate the fraction of binding sites observed.

RNA electrophoretic mobility shift assays. The following sequences from each of the four beta-actin sites were cloned in the pGEM-T vector:

site #1 = 5'-TCGCCTTAATACTTTTTTATTTTGTTTTATTTTGAATGAT-3'

site #2 = 5'-TAAAAGCCACCCCACTTCTCTCTA-3'

site #3 = 5′-CGAGGACTTTGATTGCACATTGTTGTTTTTTTAATAGTC-3′

site#4 = 5′-GGCATGGCTTTATTTGTTTTTTTTGTTTTGTTTTGGTTTTT-TTTTTTTTTTTGGCTTGA-3′

RNA probes were synthesized in in vitro transcription reactions using T7 polymerase in the presence of alpha-32P-UTP. The full length coding sequence of human ELAVL1 (HuR) was cloned into a pET vector with a GST-6xHis-Tev tag. Expression of recombinant protein from E. coli (BL21 DE3) was induced with IPTG and purified using GST columns **[133]**. REMSA assays were carried out with increasing concentrations of recombinant HuR (from 50nM to 800nM), 50mM Tris (pH 8.0), 80mM NaCl, 0.05% Triton-X, 5% glycerol, 0.5ml/ml yeast tRNA and 250pM radiolabeled RNA probe in 10ul reactions. Mixtures were incubated on ice for 30min and then resolved on non-denaturing 6% polyacrylamide TBE gels. RNA-protein complexes were visualized by phosphor-imaging.

RIP-quantitative real time PCR. HEK293 cell lysates were prepared using polysome lysis buffer as outlined in **[47]**. Likewise the lysates were used to perform RIPs as outlined in **[47]**. RNA was extracted from RIPs using TRIsure reagent (Bioline), and used to make cDNA using the iScript cDNA synthesis kit (Bio-Rad). Primer specific real time PCR (rtPCR) was performed using Platinum SYBR Green qPCR SuperMix-UDG (Thermo Fisher Scientific), and a LightCycler 1.5 instrument (Roche). The rtPCR scores

are equivalent to fold enrichments calculated using the delta-delta C_T method with the gene *COX6B1* as an internal reference **[134]**.

3.3 Results

3.3.1 DO-RIP-seq experiments are reproducible, have improved transcriptome coverage, and identify HuR binding sites

Three biological replicates of DO-RIP-seq were performed using optimized MNase conditions that would yield RNA fragments representing the binding sites of HuR. These RNA fragments were used to make cDNA libraries, which were deeply sequenced and mapped to the human genome. The first HuR DO-RIP-seq experiment yielded cDNA fragments that had an 'A' nucleotide as the first nucleotide sequenced. This resulted in a reduction in sequence quality and lower yields. Under advisement, a dark cycle run was performed on the third HuR DO-RIP-seq experiment. This resulted in improved sequence quality and yields **(data not shown)**. Also the use of UMIs increased our processed, sequenced reads approximately 2-fold when comparing the proportion of biologically duplicated sequences to unique sequences **(Fig. 3).** Note that without the UMIs we would have been unable to distinguish PCR duplicated from biologically duplicated sequences. Reads with identical sequences except in the UMI at the 3' end are considered biological duplicates and thus treated like unique sequences. On the other hand, PCR duplicates would have the same sequences and UMIs, and instead are collapsed to single reads, i.e. counted once. Without UMIs, all identical

sequences would have to be collapsed to a single read and thus reducing the number of processed reads.

HuR binding sites were previously studied using PAR-CLIP in addition to use of negative PAR-CLIP controls with the protein GFP **[35]**. Though the use of a negative control greatly improved binding site detection, its poor coverage of the transcriptome meant that the negative control was insufficient proxy of transcript abundance, and thus could not normalize the abundance of HuR binding at sites transcriptome-wide **[35]**. Therefore to quantify binding sites transcriptome-wide required that DO-RIP-seq provide sufficient coverage of the transcriptome. The transcriptomic coverage of HuR and negative DO-RIP-seq libraries were measured and compared to HuR and negative PAR-CLIP transcriptome coverages **(sub-section 3.2.2)**. This comparison revealed much higher transcriptome coverage with DO-RIP-seq libraries compared to PAR-CLIP libraries, importantly when comparing the negative libraries **(Fig. 4A)**. Further we observed good correlations between the abundance of reads at 5 nucleotide genomic intervals for HuR and negative DO-RIP-seq replicates. Importantly, the correlations among HuR DO-RIP-seq replicates or among negative DO-RIP-seq replicates were greater than correlations between HuR and negative DO-RIP-seq **(Fig. 4B)**. This indicates the specificity between transcriptomic sites bound by HuR and the quantified background.

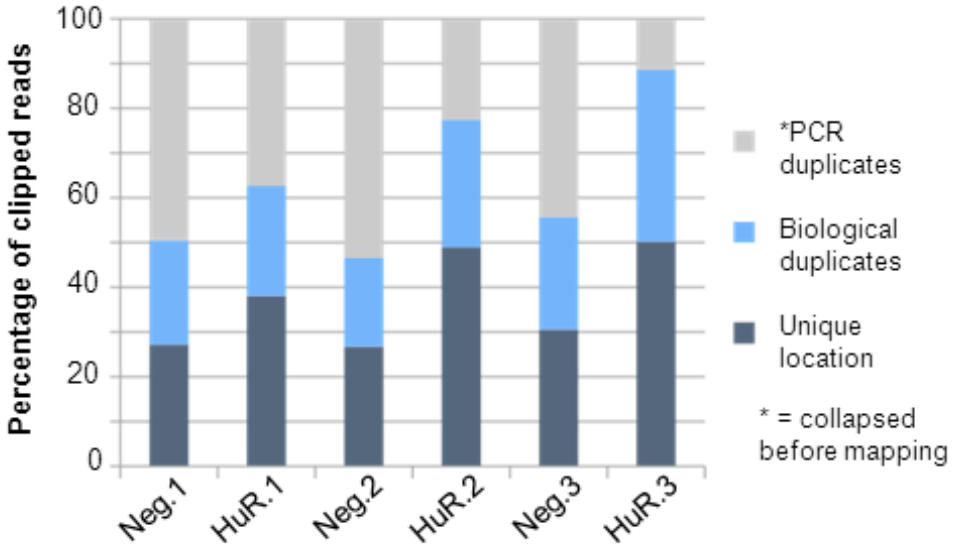

Figure 3: The bar graph depicts the percentage of clipped reads (i.e. raw reads for which an adapter could be identified and removed) for each replicate that were identified as either PCR dupllicates (same insert and same UMI), biological duplicates (same insert sequence but different UMI), or uniquely mapped reads (not duplicated). Note that in later steps PCR duplicates were collapsed to a single read before mapping.

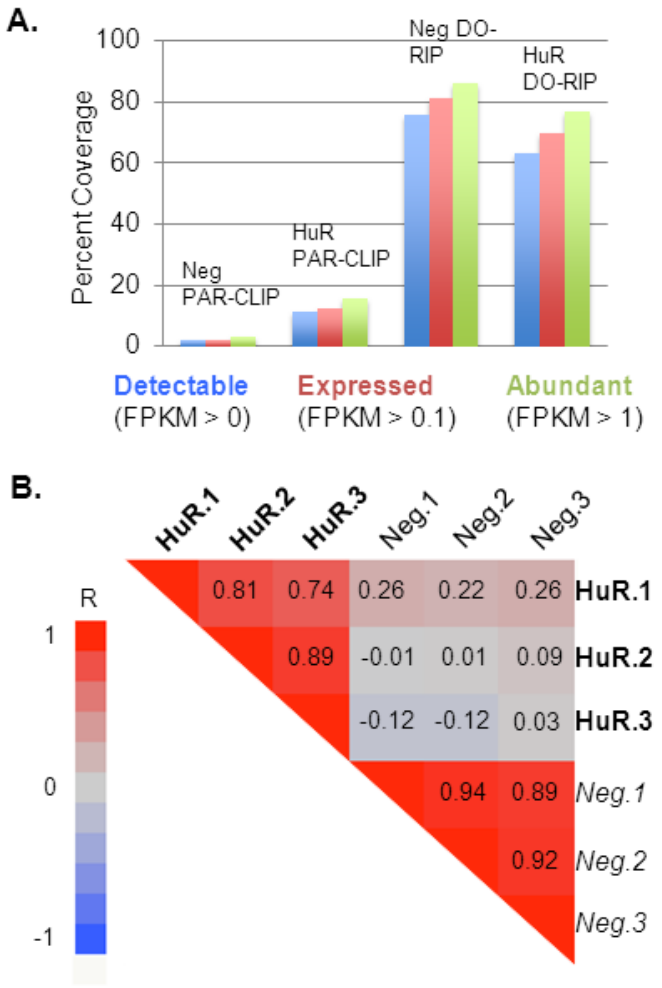

Figure 4: A. Comparison of PAR-CLIP-seq (Friedersdorf & Keene 2014) and DO-RIP-seq transcriptome coverage derived from cDNA libraries of immunoprecipitated HuR and Neg based on fragments per kilobase of exon per million mapped (FPKM). [NPC = negative control PAR-CLIP, HPC = HuR PAR-CLIP, NDR = negative control DO-RIP-seq, HDR = HuR DO-RIP-seq]. B. Comparative matrix of R-correlation values for the number of aligned reads at five nucleotide intervals of three replicates of HuR (HuR.1/.2/.3) and negative (Neg.1/.2/.3) DO-RIP-seq libraries.

3.3.2 DO-RIP-seq reproducibly identifies a greater number of functional HuR binding sites

Next HuR binding sites were quantified by calculating local enrichment values of HuR-bound reads over read values from the corresponding negative IP at five nucleotide intervals across the transcriptome. A histogram of these enrichment values yielded a bimodal distribution similar to that observed for mRNA enriched in prior RIP-chip studies **(Fig. 5A) [24]**. Similar to RIP-chip, Gaussian mixture modeling (GMM) was applied to distinguish the enriched set. GMM was used to calculate the probability of association for each five nucleotide bin using log of odds (LOD) scores **[24]**. Then p-values were calculated to determine five nucleotide bins that were reproducibly enriched for HuR reads versus negative reads among the biological replicates. Finally, to identify binding sites we selected bins with LOD values greater than zero (i.e. a greater probability of belonging to the HuR mixture than the negative mixture) and p-values less than 0.05 and if three or more consecutive five nucleotide bins met these criteria the bins were merged to define the binding site. This yielded a total of 373,845 reproducible transcriptome binding sites for HuR in HEK293 cells **(Fig. 5B)**. These included many previously reported binding sites of HuR such as in CCND1 3′UTR **(Fig. 6A)**, p27/CDKN1B 5′UTR **(Fig. 6B)**, p21/CDKN1A 3′UTR **(Fig. 7A)**, and MYC 3′UTR **(Fig. 7B) [73, 135-137]**. These sites were also identified previously using traditional biochemical and molecular biology techniques. Some of these targets (e.g. CCND1, and p27) were not reproduced in all HuR PAR-CLIP studies, however all were reproduced by DO-RIP-seq.

Also these sites had much lower read depth (measured in reads per million, RPM) in PAR-CLIP compared to DO-RIP-seq in most cases. This suggests that the greater coverage offered by DO-RIP-seq allows for greater detection and reproducibility. Additionally, we successfully performed DO-RIP-seq experiments for HuR in mouse cells and identified many of the same sites in mouse as we observed in human **[138]**. While producing more HuR binding sites than reported in any other study, the locations and nucleic acid sequence characteristics were similar to previous observations **[54, 73, 79, 139, 140]**. For example, the majority of HuR binding sites were in introns and 3'UTRs of mRNA **(Fig. 8A)**, and U-rich motifs were most frequently observed in the binding sites **(Fig. 8B)**.

Next we sought to demonstrate that DO-RIP-seq identifies functional sites and evaluate how DO-RIP-seq identified HuR targets compared to PAR-CLIP identified HuR targets. To do this we compared HuR mRNA targets identified by DO-RIP-seq to changes in mRNA abundance following HuR siRNA knockdown. Previously published knockdown data from the same cell lines (HEK293) were used **[54]**. Expressed mRNAs were classified as either targets identified by both DO-RIP-seq and PAR-CLIP, as targets only identified by DO-RIP-seq, as targets only identified by PAR-CLIP or as mRNAs not identified as targets by either technique (PAR-CLIP targets were taken from the same publication as the HuR siRNA data **[54]**). These groups of mRNAs were compared by cumulative distribution function plots **(Fig. 8C)**. Since HuR is proposed to generally

stabilize target mRNAs we expected to see a rightward shift in the plot, thus representing greater mRNA abundance in the control knockdown versus the HuR siRNA knockdown. It was evident from this analysis that HuR targets identified by both techniques showed the greatest rightward shift with targets identified by DO-RIP-seq only following closely in second. HuR targets identified by PAR-CLIP only showed a modest but significant difference from mRNAs not identified as targets by either technique. This result shows than in general HuR targets identified by DO-RIP-seq are more functionally responsive to siRNA knockdown than non-targets. Furthermore, it appears that DO-RIP-seq is able to identify additional functionally responsive target mRNAs that are missed by PAR-CLIP.

Together, this demonstrates that DO-RIP-seq can identify transcriptome-wide, functional HuR binding sites and that these sites are similar to those previously reported with the notable distinction that DO-RIP-seq has identified many more functional HuR binding sites.

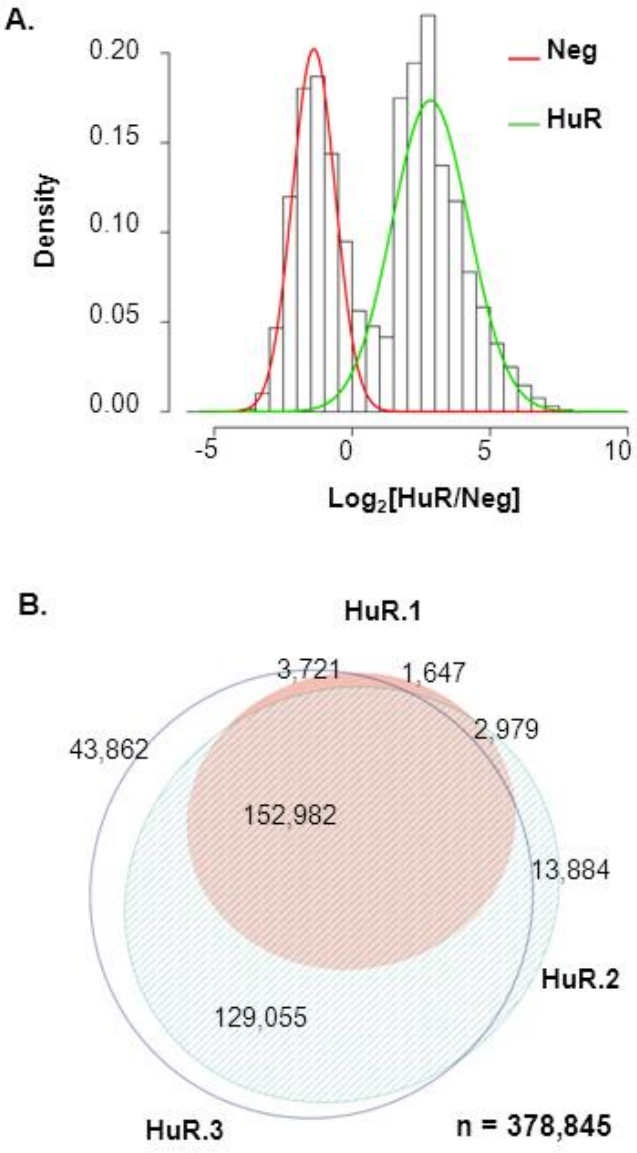

Figure 5: A. Histogram of log2 transformed average read enrichment (HuR IP over Nep IP) for each 5 nucleotide bin, with a characteristic bimodal pattern of enrichment. Mixture modeling was performed to determine the probability of each 5 nucleotide bin belonging the distribution on th right (enriched in HuR IP relative to Neg IP) versus belonging to the distribution on the left (depleted in HuR IP). The mixture model of the HuR enriched sites is in green and the mixture model of the Neg enriched sites (HuR depleted) is in red. B. Venn diagrams for binding sites independently identified in each HuR DO-RIP-seq replicate (HuR.1/.2/.3).

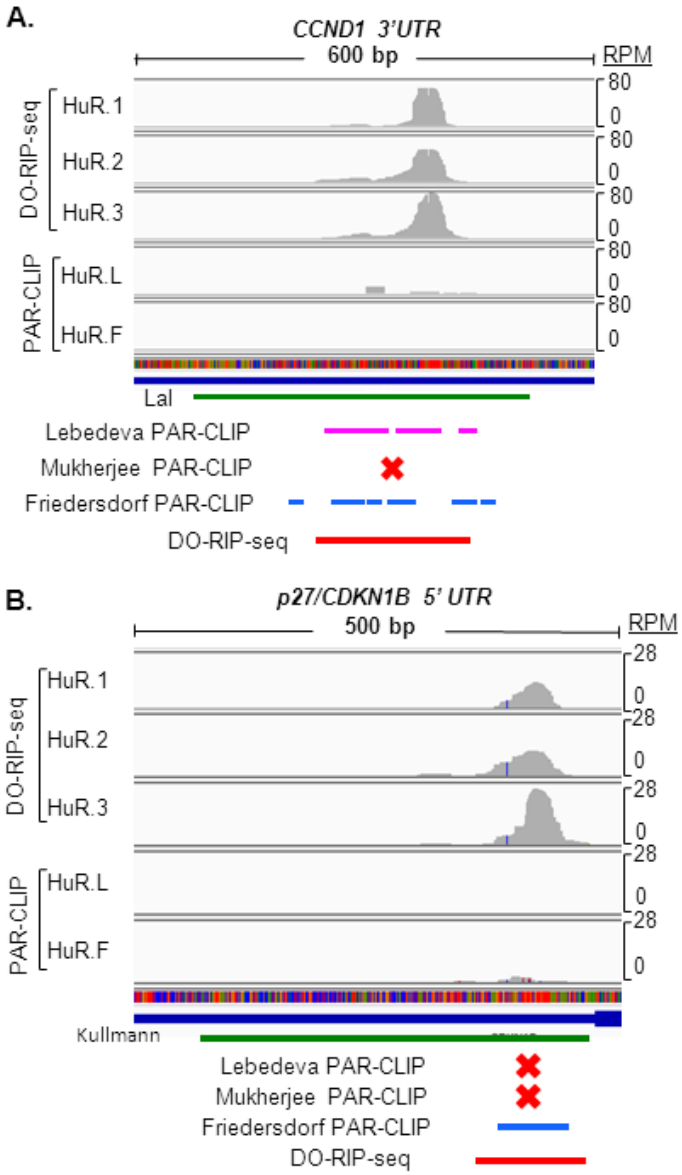

Figure 6: IGV browser snapshots with coverage tracks for three biological replicates of HuR DO-RIP-seq (HuR.1/.2/.3), HuR PAR-CLIP by Lebedeva et al. 2011 (HuR.L), and HuR PAR-CLIP by Friedersdorf & Keene 2014 (HuR.F). The green bars represent experimentally validated HuR binding sites from previous studies (Lal et al. 2004, Kullmann et al. 2002). Magenta bars represent binding sites from HuR PAR-CLIP by Lebedeva et al. 2011. Orange bars are binding sites from HuR PAR-CLIP by Mukherjee et al. 2011. Light blue bars are from HuR PAR-CLIP by Friedersdorf & Keene 2014. Red bars are HuR DO-RIP-seq binding sites. Bars with red "X" depict binding sites not observed in a PAR-CLIP study. (A) CCND1 3'UTR binding sites. (B)

p27/CDKN1B 5′UTR. Note that this site is present in DO-RIP but only in one of the three PAR-CLIPs. Also note the low coverage of this site in PAR-CLIP compared to DO-RIP-seq.

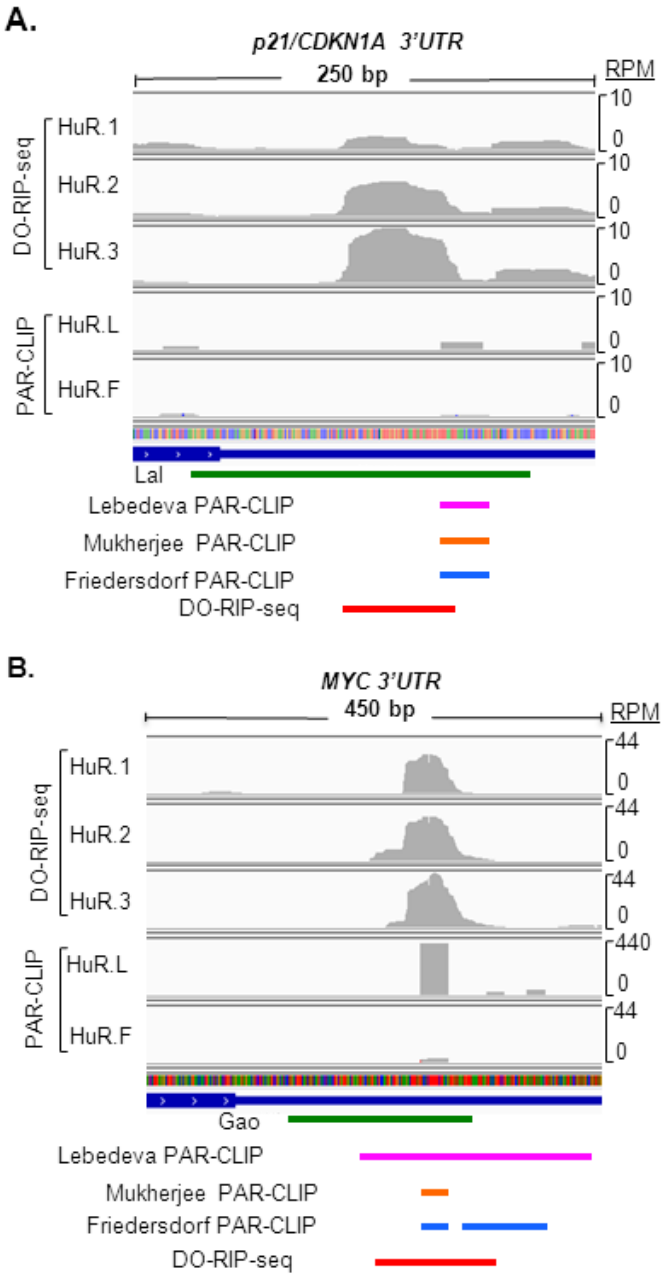

Figure 7: IGV screenshots of previously reported HuR binding sites (Green bars, Gao & Keene 1996, Lal et al. 2004). The color scheme is the same as in Figure 6. A. p21/CDKN1A 3'UTR binding site which is present in all HuR binding studies. B. HuR site in MYC 3'UTR which was observed in all studies.

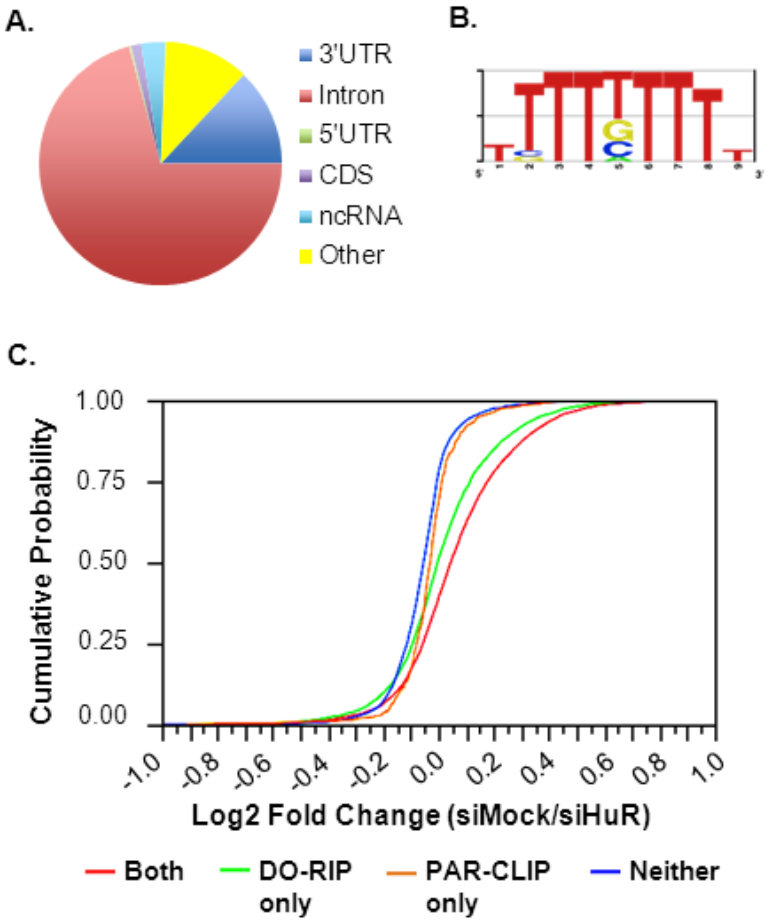

Figure 8: (A) A comparison of the abundance of HuR binding sites in locations across the transcriptome [3'UTR = 3' untranslated region, 5'UTR = 5' untranslated region, CDS = coding sequence exons, ncRNA = non-coding RNA. (B) Sequence logo of the most frequently observed cDNA nonamer in HuR binding sites. (C) CDF plot comparing log2 fold-change in mRNA expression following HuR siRNA knockdown for HuR targets identified by DO-RIP-seq, PAR-CLIP, both, or neither. A rightward shift in the curve suggests a greater proportion of targets proportion of targets since HuR is proposed to generally stabilize mRNAs. Targets identified by both techniques were most responsive to HuR knockdown. Targets identified by DO-RIP-seq only were more responsive than ones detected by only PAR-CLIP or by neither technique.

3.3.3 Saturation of HuR binding sites definitively shows that HuR binds to unstructured sequences

The increased number of binding sites for HuR directed me to test whether a complete set of HuR binding sites was identified. A saturation analysis was performed to compare the percentage of binding sites identified when randomly sampling proportions (10%, 30%, 50%, 70%, and 90%) of aligned sequence reads from our DO-RIP-seq libraries **(Fig. 9A)**. Reads were sampled from each of the three replicates and the binding sites calculated sites using the same binning and mixture model approach with the same criteria as detailed for the full datasets. Extrapolating from the saturation curve it was evident that additional sequencing would not produce proportionally more HuR sites in HEK293 cells. This is especially pronounced for sites with higher LOD scores. This saturation analysis suggests that we have identified nearly all highly probable HuR binding sites and a large proportion of all HuR binding sites **(Fig. 9A & 9B)** in the given cell line and condition.

Another use of DO-RIP-seq data is for identifying features surrounding sequence motifs that determine binding, for example the presence or absence of secondary structure. In order to identify structural features of bound sites the ability to identify "non-bound" HuR sites is required as well. Given that HuR is a sequence specific RBP, a non-bound site is characterized as one that contains an expressed U-rich motif, but is not detected as a binding site in the DO-RIP-seq data. Another interpretation of "non-bound" would be that they were bound sites not detected due to insufficient sequencing.

However, this is less likely to be the case because of the empirical evidence suggesting that deeper sequencing would not have yielded more binding sites in proportion to the sequencing depth **(Fig. 9A)**. To determine whether local secondary structure precludes binding by HuR, a HuR motif was defined from DO-RIP-seq binding sites using a k-mer approach **(Fig. 8B)**. Next, expressed 3'UTRs were searched for this motif, and each detected motif classified as either bound (LOD > 0) or non-bound (LOD < 0). To calculate the probability of the motifs occurring in localized structured regions we examined each binding site plus 150 nucleotides of surrounding sequence and calculated site accessibility using the RNAplfold program **[141]**. Finally, "area under the receiver operating characteristic analysis" was used (ROCR package in R) to evaluate how well local site accessibility predicted binding. As shown in **Fig. 10** site accessibility was a stronger predictor of HuR binding when using DO-RIP-seq than for our previous HuR PAR-CLIP data **[35]**. This finding corroborates previous studies in concluding that HuR binds to specific sequences in unstructured regions of RNA **[76, 78]**.

Taken together, these data show that increased sequence coverage allows for the unique ability of DO-RIP-seq to identify most of the high probability binding sites, and thus, also identify meaningful non-bound sites as demonstrated by the presence of local secondary structure surrounding non-bound sites. Also, it demonstrates that DO-RIP-seq preserves the local binding preferences for an RBP in the cell because we detected

expressed, non-bound sites containing HuR motifs, and these were predominantly in structured regions.

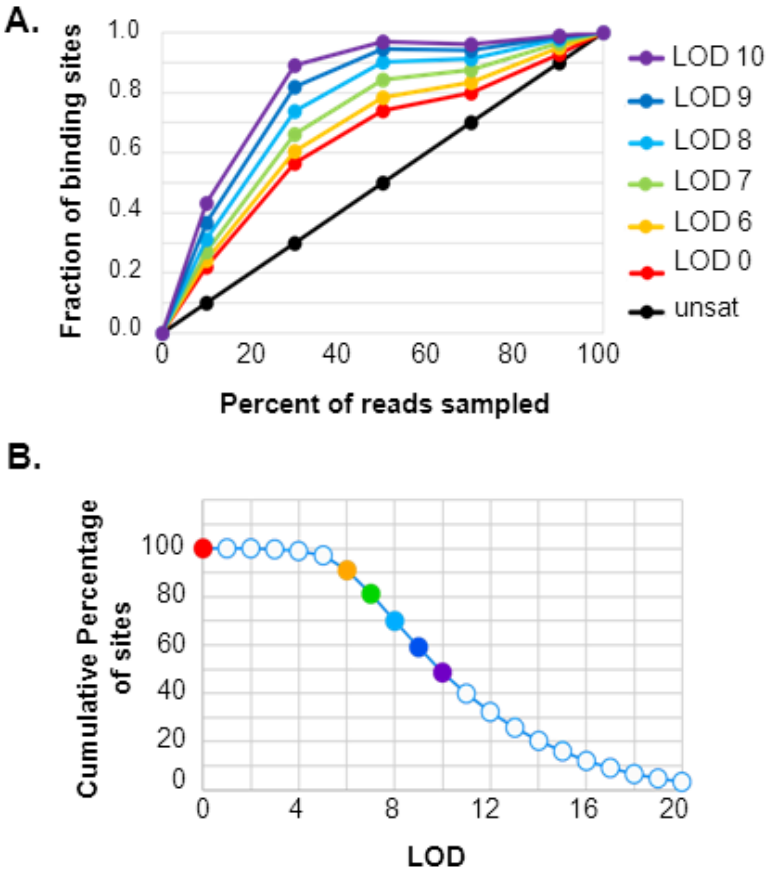

Figure 9: (A) HuR binding site saturation analysis. The fractions of significantly enriched (p value < 0.05) HuR binding sites were plotted for different percentages (10%, 30%, 50%, 70%, 90%) of random sampled HuR DO-RIP reads. Black line indicates expected trend for completely unsaturated set. Colored points indicates fraction of binding sites with LOD scores greater than the associated value. (B) Cumulative percentage of all HuR binding sites for sites with LOD scores greater than the associated value. Colors match colors from panel *A* to indicate the cumulative percentage for each saturation curve.

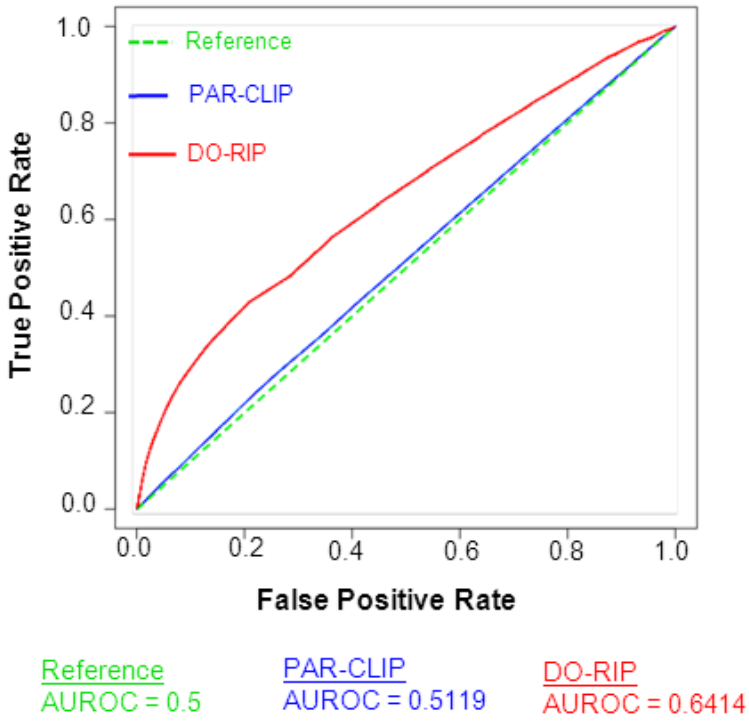

Figure 10: Area under the receiver operator characteristic (AUROC) analysis testing how well site accessibility predicts binding by HuR comparing DO-RIP-seq or PAR-CLIP data. Red line is DO-RIP-seq, blue line is PAR-CLIP, and dashed green line is x = y. Greater AUROC = strength of prediction.

3.3.4 DO-RIP-seq enrichment scores correlate with binding strength

Since the LOD scores represent binding site enrichment we decided to test the relationship between LOD score and binding strength. As an example, we selected four distinct sites within ACTB/β-actin mRNA that has a wide range of LOD scores **(Fig. 11A)**. Two of the selected sites (site #1 and #4) have high LOD values, another (site #3) falls below the median LOD score, and the final site (site #2) being the lowest of the four, with a LOD score in the bottom 1.5% of all global HuR binding sites. In addition, HuR's binding to site #4 was previously shown to result in stabilization of the mRNA **[142]**. For

comparison, REMSA was performed with various amounts of HuR protein, incubated with radioactive probes from each site, and analyzed on non-denaturing polyacrylamide gels. A strong correlation between the DO-RIP-seq LOD scores and apparent mobility shifts of the corresponding probes was evident **(Fig. 11B)**. This experiment suggested that LOD scores of transcriptome-wide binding sites quantified using DO-RIP-seq directly reflect relative differences in binding strength at specific RNA sites in the cell, although tests of additional mRNAs will be needed to generalize this conclusion.

Assuming that LOD scores indicate the binding strength of HuR, certain specific sequence characteristics associated with HuR binding would be expected to demonstrate higher scores. To test this assumption, we analyzed HuR DO-RIP-seq binding sites for trends between the frequencies of different U-rich sub-motifs and LOD scores. Potential LOD score cutoffs producing five binding site groups that were each approximately 20% of all sites were selected and all possible 7-mers enumerated. 7-mers that were similar (the sub-motifs) were grouped and the frequencies of their occurrences among the LOD score groups were calculated. The result was that AU-rich sub-motifs had increasing frequencies at higher LOD scores, while CU- and GU-rich sub-motifs had decreasing frequencies at higher LOD scores **(Fig. 12A)**. Note that each LOD score group contained some proportion of each of the U-rich sub-motifs, and that many of our binding sites contained combinations of these sub-motifs. This raised the question as to how sub-motif combinations relate to LOD scores for binding sites. We hypothesized that binding

sites with combinations of preferential motifs for HuR binding would have higher LOD scores. We tested this hypothesis using RNAcompete Z-scores of 7-mer motifs for HuR, where a Z-score is an empirical value that measures the preference of recombinant HuR for the 7-mer in an *in vitro* reaction **[32]**. A comparison of the summation of 7-mer Z-scores at binding sites and the DO-RIP-seq LOD score yielded a good correlation; Pearson's R = 0.4996 **(Fig. 12B)**. Therefore, we conclude that HuR binding sites with high LOD scores contain combinations of preferential sequence motifs. However, it should also be noted that while there is a good correlation between DO-RIP-seq LOD scores and *in vitro* binding measurements, the relationship is not perfect. We believe that at least some of these differences can be attributed to the differences between HuR binding under cellular conditions (DO-RIP-seq) versus *in vitro* binding (RNAcompete). Taken together, we have quantitatively identified cellular binding sites of HuR to distinguish sites of strong interactions that are partially based on frequencies of particular RNA sub-motifs.

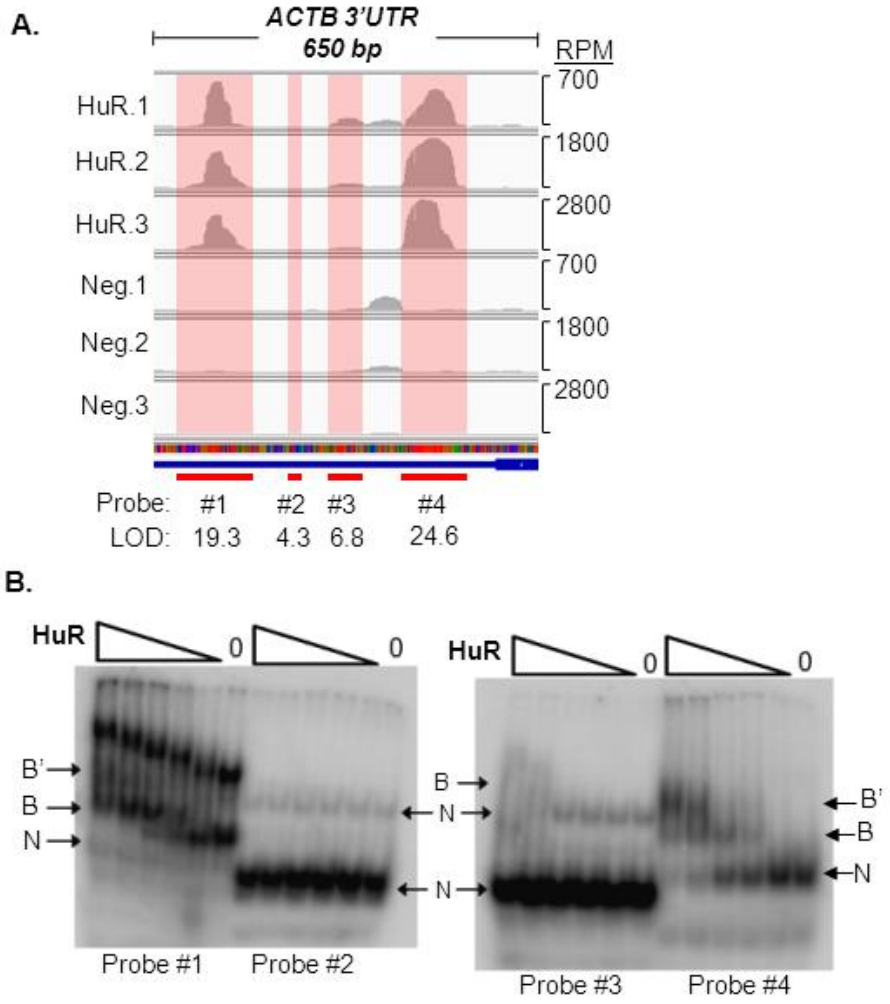

Figure 11: (A) Comparison of 4 HuR binding sites detected in the 3'UTR of *ACTB* mRNA. The read depth of the sites are displayed reads per million (RPM) and the calculated LOD scores for each site are depicted. Note that each binding site shown in red has a corresponding LOD score that suggest sites 1 and 4 have the highest probability of binding. (B) Images of RNA electrophoretic mobility shifts performed by Matthew Friedersdorf. Radiolabeled RNA probes representing each of the detected binding sites in *ACTB* mRNA were used. Note that the degree of each shifted band (HuR-probe complex) visually correlates with the LOD score for that site. B = bound, B' = bound, second shift, N= non-bound.

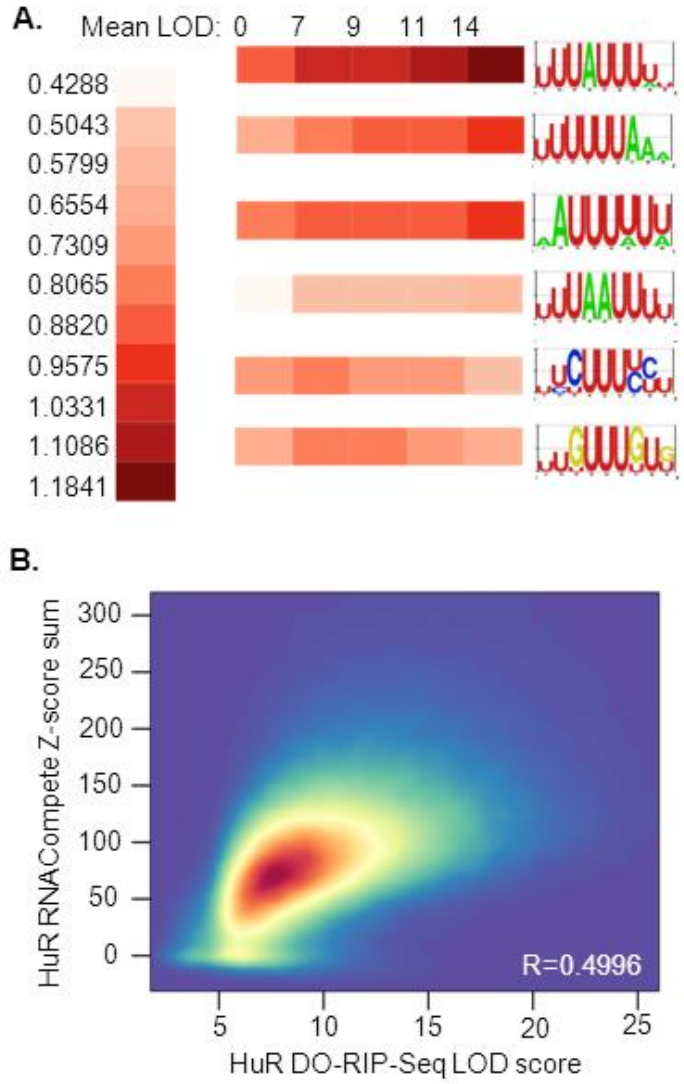

Figure 12: (A) A series of U-rich binding motifs and their frequencies of enrichment in HuR binding sites, grouped by LOD score. Note that the direction of increasing motif frequency with increasing LOD score reverses with the two bottom motifs. (B) A correlation heat map between HuR DO-RIP-seq binding site LOD scores and the sum of Z-scores from HuR 7-mer motifs enriched in RNAcompete. Pearson's R = 0.4996. Colors from blue to red represent increasing density of HuR binding sites. The coordinates with greatest density of binding sites contribute the most to the R correlation value. RNAcompete data is from Ray et al. 2009.

3.3.5 DO-RIP-seq identifies whole transcript targets of HuR and enriched functional gene sets

RNA binding proteins bind to local sites within mRNA or pre-mRNA and these local binding events ultimately decide the fate of the whole transcript. While whole transcripts can be defined as RNA targets based on whether they contain a single mRNA binding site, it is more robust to define RNA targets by quantifying binding of the whole transcript by the ratio of RBP binding relative to a negative control **[20, 24, 143]**. The demonstration that negative DO-RIP-seq controls have good coverage of the transcriptome led us to believe that the negative background binding could be used to quantify whole mRNA targets as is done with RIP-chip experiments when used to normalize transcript abundance. Therefore, whole transcript targets of HuR were quantified by summing up all reads in each mature transcript, including reads at binding sites as well as background reads, and normalizing the abundance of reads in the HuR DO-RIP-seq against the NEG DO-RIP-seq. This produced a distribution of enrichment scores for transcripts that we call RIP-Seq-Like scores (RSL score, **Fig. 13A**). Interestingly, the distribution of RSL scores appears to be bimodal similar to the bimodal distributions seen in RIP-chip enrichment values. Furthermore, the depleted distribution (left most) consists mostly of transcripts that do not have any DO-RIP-seq (not shown) or PAR-CLIP HuR binding sites **(Fig. 13A)**, while the enriched distribution (right most) consists of transcripts that contain many DO-RIP-seq or PAR-CLIP HuR binding sites.

To evaluate whether RSL scores are quantitatively similar to enrichments in undigested, whole transcript RIPs, we performed HuR RIPs followed by quantitative real time PCR (RIP-rtPCR) for 28 randomly selected transcripts representing the whole range of RSL scores (from the most enriched to the most depleted). We observed strong correlations between RSL scores and RIP-rtPCR scores with a Pearson's R = 0.72 **(Fig. 13B)**. Also many of these targets displaying low enrichments (bottom-left of the graph) were ones that had no PAR-CLIP binding sites **(Fig. 13B)** showing the qualitative agreement between the methods. In addition, we compared RSL scores to HuR RIP-chip derived LOD scores from our previous experiments **[54]**.

The LOD scores used in RIP-chip are analogous to LOD scores from DO-RIP-seq but with the distinction that RIP-chip LOD is a probabilistic measure of whole transcript enrichment rather than binding site enrichment. We compared the RSL values to the RIP-chip LOD scores by using a cumulative distribution function of RSL scores for groupings of increasing LOD scores. This revealed a strong relationship between RSL and RIP-chip LOD with increasing LOD score groups having higher RSL scores **(Fig. 14)**. This indicates that the range of whole transcript enrichments we observed using DO-RIP-seq is similar to those found by our RIP-chip procedure. This observation suggests that the sum of all binding sites from digested RIPs is quantitatively very similar to enrichments seen for whole transcripts and that each site proportionally affects the targeting of the whole transcript. However, as demonstrated below, identification of

specific binding sites, as well as whole transcripts, is exceedingly important for understanding the impact of combinatorial binding factors on the functions of each mRNA species.

To determine whether whole transcript enrichments from DO-RIP-seq experiments can identify functional gene sets, we performed a Gene Set Enrichment Analysis using RSL analysis data. Similar to previous observations from HuR RIP-chip **[24]**, we observed that HuR targets were significantly enriched for many predicted microRNA (miR) target gene sets **(Fig. 15A)**. While the majority of predicted miR target gene sets were enriched, there were some miR target gene sets that were not significantly enriched.

Given that HuR can antagonize AGO-miR functions when bound close to AGO-miR binding sites **[54, 87]**, we used our binding site data to test whether HuR binds within 10 nucleotides of a miR site at frequencies that were greater than that of a shuffled miR set. We selected miR-21, miR-155, and miR-17-5p as candidates whose targets were significantly enriched in HuR DO-RIP-seq, and miR-185, miR-326, and miR-491 as candidates having no significant enrichment in HuR DO-RIP-seq. We observed that HuR bound near the enriched miR seed sites (miR-21, miR-155, and miR-17-5p) at frequencies greater than that of the shuffled miRs, i.e. greater than random **(Fig. 15B)**. However, the frequencies for HuR binding close to the non-enriched miR seed sites (miR-185, miR-326, and miR-491) were approximately the same as shuffled miR sites

(Fig. 15B). Given that HuR has been found to generally stabilize mRNA targets while AGO-miR binding is associated with decay of targeted mRNAs, it has been proposed that antagonism of AGO function by HuR may control the turnover of HuR mRNA targets **[54, 87]**. These results support the conclusion that whole transcript target analysis using DO-RIP-seq can reveal functionally related RNA targets, and together with binding site information, illuminate the importance of combinatorial relationships among RNA-binding proteins at the binding site level.

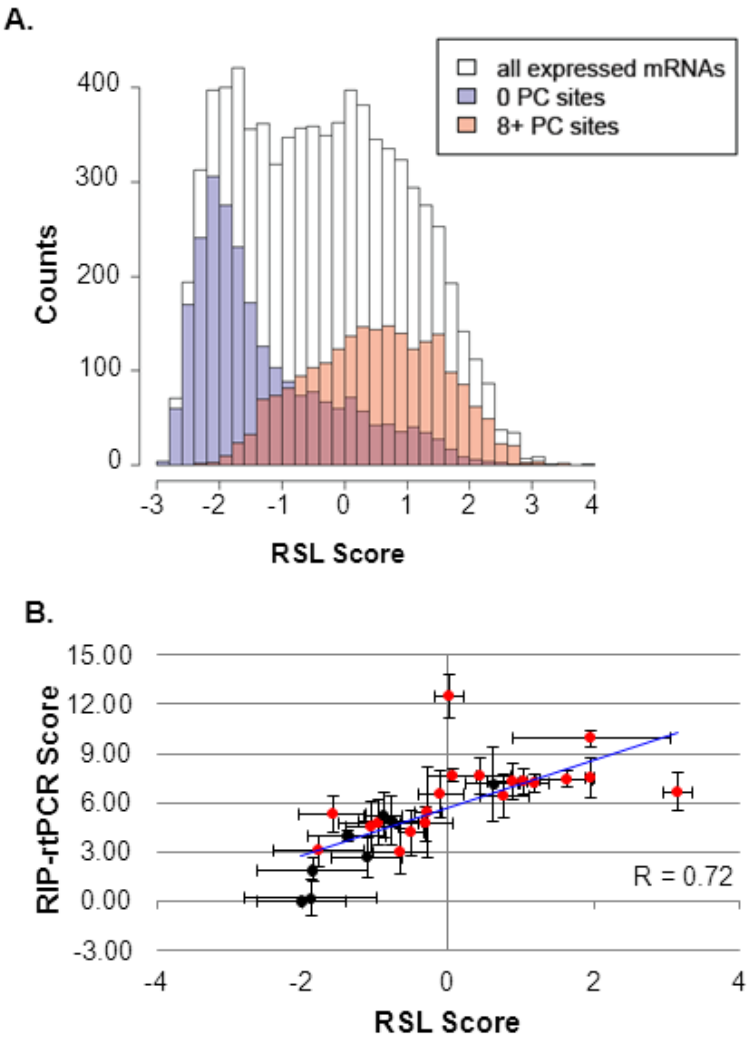

Figure 13: (A) RSL score distrribution for mRNA expressed in HEK293. The distribution for transcripts having 0 HuR PAR-CLIP (0 PC) sites and 8 or more (8+) PC sites is merged with the RSL distribution. HuR PAR-CLIp data was from Mukherjee et al. 2011. Note the separation of the distributions for the 0 PC sites and the 8+ PC sites. (B) Correlation between RSL score and RIP-rtPCR enrichment score for mRNA representing RSL score distribution is strong (Pearson's R = 0.72). Red points are genes with HuR sites according to PAR-CLIP, and black points are genes with no HuR PAR-CLIP sites.

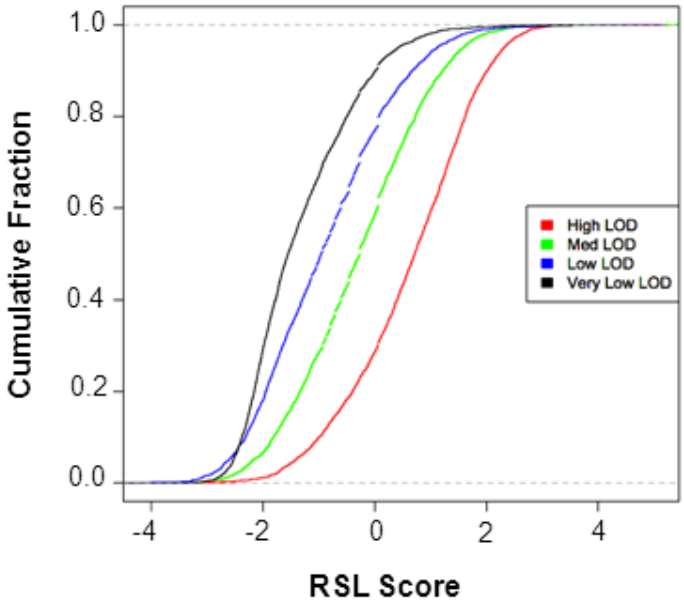

Figure 14: DO-RIP-seq RSL score and RIP-chip LOD score for whole transcript association agree. Cumulative distribution functions displaying the RSL score distribution for genes grouped by the HuR RIP-chip LOD score. Red, high RIP-chip LOD; green, medium RIP-chip LOD; blue, low RIP-chip LOD; black, very low RIP-chip LOD.

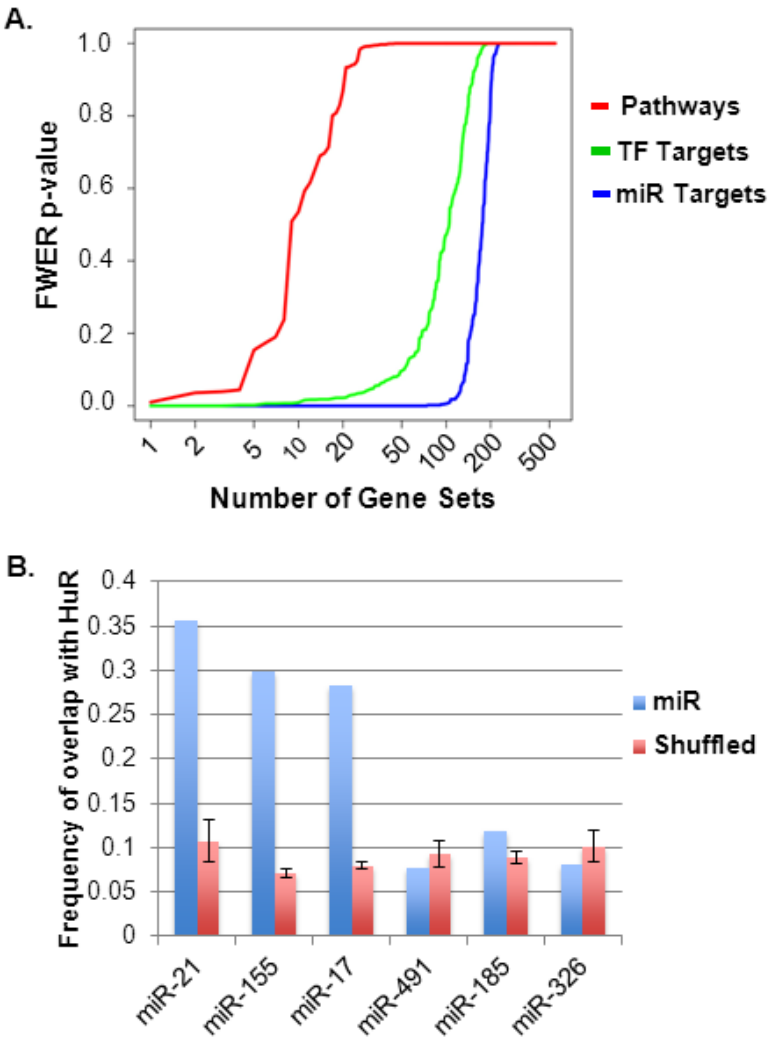

Figure 15: (A) A comparison of the number of gene sets obtained from GSEA using HuR DO-RIP-seq mRNA target analysis. Note the large number of significantly enriched microRNA targets gene sets (miR targets, blue) compared to transcription factor target gene sets (TF targets), and targets in canonical pathways (Pathways, red) based on their FWER (family wise error rate) p values of the normalized enrichments. FWER < 0.1 is considered significant. (B) The frequency of overlap between predicted miR sites and HuR binding sites compared with the frequency of overlap between predicted miR sites shuffled on 3'UTRs and HuR binding sites from DO-RIP-seq data. The miRs shown are representative of those having gene sets significantly enriched (miR-21, miR-155, miR-17), and not significantly enriched (miR-326, miR-185, miR-491) in HuR DO-RIP-seq mRNA target score analysis.

3.4 Discussion

3.4.1 Quantification of protein-RNA interactions with DO-RIP-seq in comparison to other procedures

Among the goals of this work was to quantify cellular RBP binding sites using our traditional RIP method, and to relate quantitative binding site values to that of whole transcript targeting. On the other hand, one of the challenges of quantifying cellular binding using CLIP procedures has been the necessity to consider RNA abundance, because unlike DNA that uses chromatin immunoprecipitation (ChIP) procedures, each RNA copy number per cell is variable. Thus, one must account for the variations in RNA abundance in order to assess quantitative relationships among RNA targets of a RBP. However, RIP techniques have consistently addressed this problem by measuring background levels based on a negative IP control. This allows one to calculate enrichment by normalizing transcript abundance in the positive IP to the background IP, thus accounting for differences in RNA abundance **[24, 117, 143, 144]**. In developing DO-RIP-seq we went further by normalizing local regions against background to quantitatively determine binding site enrichment. It is possible that some negative IP's may introduce biases such as the binding of aptameric RNAs to antibodies. However, we have found that our antibody-based approach to identify and quantify binding was complemented by non-antibody based Delta-SHAPE experiments determining RBP binding sites in XIST lincRNA **[138]**. This suggests that antibody-based biases have a negligible effect on DO-RIP-seq binding sites. Nevertheless, as an

alternative to negative IPs, input RNA could also be utilized by making libraries taken from the same RNase-treated lysate used in the IP and prepared using the same cDNA library protocol. We have found that aligned reads from negative IPs correlate well with "input" RNA making inputs a suitable alternative (data not shown).

Although at its inception CLIP techniques did not attend to background **[25, 30, 31, 120],** it has since been shown that significant background remains in CLIP experiments **[35]**, and that accounting for it greatly improves identification of binding sites **[36-41]**. However, it is not known whether this is due to improved signal-to-noise detection or whether the difference in enrichment represents quantitative differences in binding to various sub-motifs **[33, 38, 145]**. Some CLIP experiments have interpreted binding measurements as quantitative. However, those experiments did not formally demonstrated a direct relationship between CLIP enrichment and binding strength **[34]**. Here we have shown that by comparing local enrichment of binding sites over negative IPs we can generate a continuously distributed and quantitative value for each binding site, i.e. a LOD score. We have shown that the values of these LOD scores are directly related to *in vitro* binding strength.

Several *in vitro* methods have been developed recently to quantify binding to short stretches of RNA *in vitro* **[146]**. For example, RNA Bind-n-Seq calculates dissociation constants *in vitro* for canonical and non-canonical motifs that can be bound by an RBP, and has been used to complement CLIP in order to distinguish functional

sites from non-functional sites in CLIP data **[147]**. While DO-RIP-seq LOD scores show a relationship with *in vitro* binding strength they are not dissociation constants, as we could not incrementally vary the amount of HuR in the cell in our experiments. However, DO-RIP-seq enrichment values represent relative binding strength as it occurs in the cell. Importantly, we showed that HuR LOD scores are the result of the combination of multiple preferential motifs within the site. Furthermore, our HuR binding sites displayed a consistent trend between the frequency of optimal sub-motifs and LOD scores, suggesting that in cells HuR may have a slight preference for particular sub-motifs. Since other *in vitro* studies suggest that there is no preference for HuR binding to these sub-motifs this may reflect cellular combinatorial interactions with other RNA-binding proteins.

3.4.2 The importance of RNA binding site saturation

The number of HuR binding sites identified with DO-RIP-seq approached full transcriptome saturation for sites with higher LOD scores. We also demonstrated that higher LOD scores correlate well with strength of binding. Together, this indicates that deeper sequencing would not yield proportionally more of these stronger binding sites, and thus, we were able to identify nearly all strong, DO-RIP-seq detectable HuR binding sites in HEK293 cells. We believe the reason DO-RIP-seq can achieve binding site saturation is the relatively high efficiency immunoprecipitation of native RNPs versus using UV-crosslinked immunoprecipitations. This improved efficiency is evident when

comparing the number of PCR cycles needed for DO-RIP-seq versus various CLIP protocols; with DO-RIP-seq requiring between 11 to 14 cycles while lower efficiency CLIP typically requires cycle numbers in the mid to upper 20s **[41]**. However, it should be noted that even with the improved efficiency of DO-RIP-seq, that we may not have been able to saturate all binding sites, especially the weak and possibly transient HuR-RNA interactions. Furthermore, some classes of binding sites were more reticent to saturation; for example, intronic binding sites were less saturated than exonic sites because of the lower abundance and higher turnover of introns relative to exons. Nevertheless, we showed how this saturation of DO-RIP-seq binding sites allowed better discrimination of bound versus non-bound motifs as evidenced in the predictive relationship between binding and lack of RNA secondary structure **(Fig. 10)**, and in uncovering additional functionally responsive target mRNAs **(Fig. 8C)**.

We have only tested for saturation with HuR so we cannot say for certain whether other RBPs will show similar completeness. However, given HuR's pervasive binding across the transcriptome it is reasonable to believe that other RBPs will be similar or better than HuR in terms of saturation. Furthermore, the first two sequence runs were under-sequenced (26M and 47M uniquely mapped reads) compared to the third replicate (80M uniquely mapped reads) due to an artifact affecting sequencing depth as explained in Materials & Methods. This was addressed in the third run and

resulted in significantly more reads. Thus future runs should lead to greater sequence depth and greater overall saturation.

We believe that the ability of DO-RIP-seq to accurately distinguish bound motifs from non-bound motifs will pave the way for further discovery of more subtle sequence characteristics that influence RBP binding in cells. For example, it may be possible to find spatial relationships between bound and non-bound motifs of two different RBPs to uncover cooperative or competitive interactions between the RBPs in cells. Also, we suggest that saturation will be useful when studying RBP binding during dynamic biological processes or responses to perturbations. Dynamic conditions can modulate the binding of a RBP resulting in partial to complete loss or gain in binding. Moreover, these dynamic conditions affect transcript abundance. By saturating the binding site detection one has more confidence that changes in binding observed at specific sites are truly due to a biological rather than a technical feature. In addition to accounting for transcript abundance using a negative control IP (or matched input), and being able to quantify binding using scores like a LOD, we expect that DO-RIP-seq experiments can robustly detect and quantify biologically significant binding site changes in cells and tissues.

3.4.3 Utility of whole-transcript analysis

An unexpected but gratifying application of DO-RIP-seq is the ability to quantify RNA binding sites as well as whole-transcript targets of HuR. In this study, we found

that the normalized aggregate of the enrichments of all binding sites across the transcripts (i.e. RSL scores) is proportional to the enrichment of whole, undigested transcripts **(Fig. 13B, 14)**. While knowing the sites of RBP binding may be essential for understanding the mechanisms of regulation, the functional outcome in the cell is determined on the level of the whole transcript; thus it is important to be able to relate binding sites to whole transcript binding. Moreover, having a quantitative relationship between sequence-specific binding events and the whole message makes DO-RIP-seq data amenable to integration with other globally quantifiable methods like transcriptomic profiling, metabolic labeling of RNA, ribosome profiling, and Delta-SHAPE-MaP. Although the identification of RNA binding sites using CLIP techniques has been used to define whole transcript "RNA targets", one can only speculate as to how often or how strongly these sites are bound per transcript. We think that DO-RIP-seq whole transcript target analysis provides a way to quantitatively identify functional targets while providing the sites that impact regulation of the transcript.

3.4.4 Using the DO-RIP-seq procedure to distinguish functional relationships at the whole-transcript and binding site levels

The functional interplay between HuR and miRs has been particularly interesting because they appear to have competing effects on RNA targets, namely, RNA stabilization and destabilization, respectively. Over the past several years there have been reports of HuR's antagonistic as well as synergistic effect on miR function on a single mRNA **[148, 149]**. However, evidence from transcriptome-wide studies suggests

that HuR antagonizes the destabilizing effect of miRs when bound at overlapping binding sites **[54, 87]**. We found that HuR's binding sites residing within 10 nucleotides of enriched miR target gene sets overlap predicted miR binding sites at frequencies greater than expected (**Figure** 1**5B**). This suggests that miR mRNA targets are most likely to be bound by HuR near the miR site and to antagonize AGO-miR function. However, not all predicted miR sites overlapped with HuR sites at rates greater than random. These outlying miR target sets contained transcripts that were depleted in HuR RNPs and contained significantly fewer 3'UTR binding sites. Given that HuR and miRs have roles in overlapping processes like cell cycle, apoptosis, and angiogenesis, then the functional interplay between the two is likely to be essential in these biological processes **[150, 151]**.

In summary, DO-RIP-seq is a novel methodological approach that provides quantitative scores for comparisons of whole-transcript targeting as well as in cell binding site strengths of RBPs. This approach is useful for quantifying changes in RBP binding events at the whole-transcript and binding site levels in response to environmental signals and perturbations while also allowing for more complete integration with gene regulation data. Another compelling reason to use this approach in comparison to crosslinking methods is that it preserves important biological information contained within protein-RNA interactions that evolved to be non-covalent and represent a unique biological state of gene expression. When these subtle

combinations of bonds are overridden by covalent bonds inserted by crosslinking the RNA-binding protein to the RNA, much of the biological essence is lost. Therefore, our RIP procedures preserve most of the non-covalent properties of protein-RNA interactions and enrich our ability to observe the dynamic biological changes inherent to a functional post-transcriptional outcome.

4. RBM38 is a sequence-specific RNA binding protein that extensively overlaps with HuR or TRA2B at binding sites

4.1 Introduction

I decided to investigate the transcriptome-wide binding sites of RBM38. Interestingly, the mRNA of this protein undergoes a decrease in association with ribosomes when transitioning from G1 to S phase of the cell cycle **[152]**. This is in agreement with reports showing that its induction results in G1 cell cycle arrest **[153]**. Data from *in vitro* studies using SELEX-seq and RNAcompete reveal a [GU]n motif having the highest affinity for RBM38 **[90, 93]**. However, investigations with iCLIP reveal a U-rich motif in RBM38 transcriptome-wide binding sites but did not report the [GU]n motif **[95]**. The differences between RBM38 motifs deduced from *in vitro* studies and iCLIP suggests that a more extensive transcriptome-wide investigation could reveal more about the determinants of RBM38 binding. Given the improved transcriptome coverage of DO-RIP-seq, and its non-reliance on inefficient crosslinking, I hypothesized that RBM38 binding sites identified by DO-RIP-seq would reveal RNA sequence and structural determinants of binding.

Indeed, DO-RIP-seq experiments were successfully performed on RBM38, and revealed three sequence motifs enriched in RBM38 binding sites. Like HuR, RBM38 binding sites were non-structured. Two of the three motifs detected were similar to previous reports; a [GU]n motif as reported by *in vitro* studies, and a U-rich motif as

reported by transcriptome-wide investigation with iCLIP. When RBM38 and HuR DO-RIP-seq were integrated, it was revealed that binding sites of both proteins overlapped at frequencies that were greater than random. The third motif in RBM38 sites was a purine rich motif that is known to be recognized by TRA2 proteins. DO-RIP-seq was performed on the TRA2B protein to reveal binding sites that frequently overlapped with RBM38 and these overlapping sites also contained the purine-rich motif. However, HuR and TRA2B did not exhibit significant co-binding transcriptome-wide. RSL analysis yielded continuous metrics for the whole mRNA targets of RBM38, and co-targets with HuR. Co-targets of RBM38 and HuR were enriched for regulators of biological processes such as WNT signaling, circadian rhythms, and protein ubiquitination. Therefore, I have evidence that suggests RBM38 specifically binds at locations that frequently overlaps with HuR or TRA2 proteins, and that RBM38 and HuR potentially regulates similar subsets of functionally related groups of mRNAs.

4.2 Materials & Methods

Cell lines and antibodies. The cell lines used in these studies were HEK293 (human embryonic kidney cells 293) that can be obtained from the American Type Culture Collection. HEK293 transiently expressing RBM38-3XFLAG was produced by Lipofectamine 2000 (Thermo Fisher Scientific) transfection of a plasmid containing the RBM38 coding sequence (NM017495.5) with a triple FLAG sequence in frame at the 3' end, controlled by a CMV promoter (GeneCopoeia). Antibodies used were mouse anti-

HuR (3A2) monoclonal, mouse anti-RBM38 (A8) monoclonal (Santa Cruz Biotechnology), rabbit anti-TRA2B polyclonal (Abcam), rabbit anti-HuR polyclonal (Abcam), mouse anti-FLAG M2 monoclonal (Sigma-Aldrich), HRP conjugated sheep anti-mouse IgG H+L (Jackson ImmunoResearch), HRP conjugated goat anti-rabbit IgG H+L (Jackson ImmunoResearch), normal mouse serum (Jackson ImmunoResearch), and normal rabbit serum.

Co-immunoprecipitation assays. HEK293 cells were transfected with pcRECEIVER-RBM38-3xFLAG, and harvested then lysed with CO-IP lysis buffer (10 mM HEPES, 50 mM KCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$, 0.5% NP40). Lysates were treated with 5 gel units of micrococcal nuclease per ug of total RNA for 5 minutes at 30 °C. Immunoprecipitations were performed with antibodies against FLAG peptide (for RBM38-3xFLAG), HuR, TRA2B, or with non-specific mouse or rabbit serum antibodies in CO-IP/Wash buffer (50 mM Tris-HCl pH 7.4, 50 mM NaCl, 1 mM $MgCl_2$, 0.05% NP40). Western blots were performed to detect co-immunoprecipitation of the other two RBPs; HuR and TRA2B for RBM38 IP, HuR and RBM38 for TRA2B IP, and TRA2B and RBM38 for HuR IP.

DO-RIP-seq. Wild type HEK293 were harvested and lysed using lysis byffers and procedures described in sections 2 and 3. Lysates were treated with micrococcal nuclease (New England BioLabs) to partially digest RNA, and then immunoprecipitations were performed with either antibodies specific to the RBPs, or

with non-specific serum from the same species antibodies. RNA fragments were extracted and then used to make cDNA libraries for deep sequencing on Illumina Hi-Seq 2500. Raw sequenced reads were processed to remove adapter sequences and then mapped to reference human genome hg19 using TopHat **[126]**. Mapped reads from the RBP libraries were normalized to the negative library to calculate log fold changes. Where applicable, log of odds scores (LOD), and RSL scores were calculated as previously described.

Gene Ontology (GO) Categorization. GO categories for the whole transcript targets of RBM38 only, HuR only, and both RBM38 and HuR were calculated using the gene ontology enrichment analysis and visualization tool (GOrilla, **[154, 155]**). Targets were defined as transcripts with RSL scores > 0, and all genes with $\log_2$ RPKM > 1 in the negative IPs were used as the background sets.

4.3 Results

4.3.1 RBM38 binding sites contain three sequence motifs

We investigated the transcriptome-wide binding of RBM38 using DO-RIP-Seq to identify features within its binding sites that would determine how it recognizes RNA. RBM38 DO-RIP-Seq was performed in HEK293 cells expressing endogenous levels of RBM38. Similarly to HuR, most RBM38 binding sites were in the 3'UTR and introns of mRNA **(Fig. 16A)**. Also, two additional RBM38 DO-RIP-seq experiments were performed using lysates from HEK293 transiently over-expressing RBM38-3xFLAG.

Though this lead to the detection of a greater number of binding sites, the distribution of the sites based on location remained similar to the endogenous condition **(Fig. 16B)**. Therefore, the binding sites observed as a result of over-expression were likely not detected in the endogenous DO-RIP-seq because of the low levels of RBM38 in that condition. Similarly to HuR, RBM38 binding sites were located mostly in 3'UTRs and the introns of mRNA **(Fig. 16A, 16B)**. However, RBM38 exhibited proportionally more binding to CDS compared to HuR. We also detected previously reported RBM38 binding sites like the one in the *p21* mRNA 3'UTR that overlaps the HuR binding site in this location **(Fig. 17)**. It was also reported by Cho and colleagues that RBM38 cooperates with HuR to bind *p21* mRNA in this location **[94]**.

The binding sites of RBM38 were enriched for sequence motifs; a [GU]n motif, a GU-rich motif, and a purine rich motif **(Fig. 18)**. The [GU]n and GU-rich motifs are similar to the ones reported by *in vitro* experiments and iCLIP experiments respectively **[90, 93, 95]**. However, the purine-rich motif has never been reported. Next we wanted to investigate the impact of RNA secondary structure on the ability of RBM38 to bind to its sequence motif. Our hypothesis was that sequence motifs in non-structured locations would predict RBM38 binding. We performed the same analysis that was previously described for HuR binding sites **(section 3.3.3)**, except here we used the presence or absence of the RBM38 sequence motifs to define bound or non-bound sites respectively. We observed that sites containing the sequence motifs in a non-structured region

predicted the binding of RBM38 **(Fig. 19)**. Therefore, RBM38 binds to sequence motifs in non-structured locations of RNA.

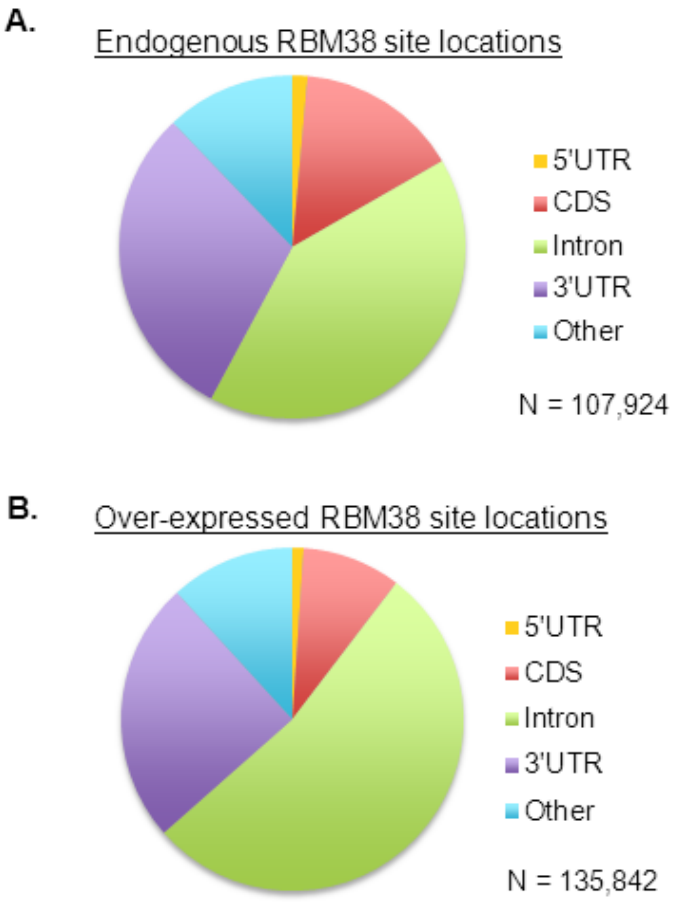

Figure 16: RBM38 binding site locations. (A) The proportions of RBM38 binding sites in transcript locations from DO-RIP-seq performed with endogenous levels of RBM38. (B) The same as a except RBM38 levels were transiently overexpressed.

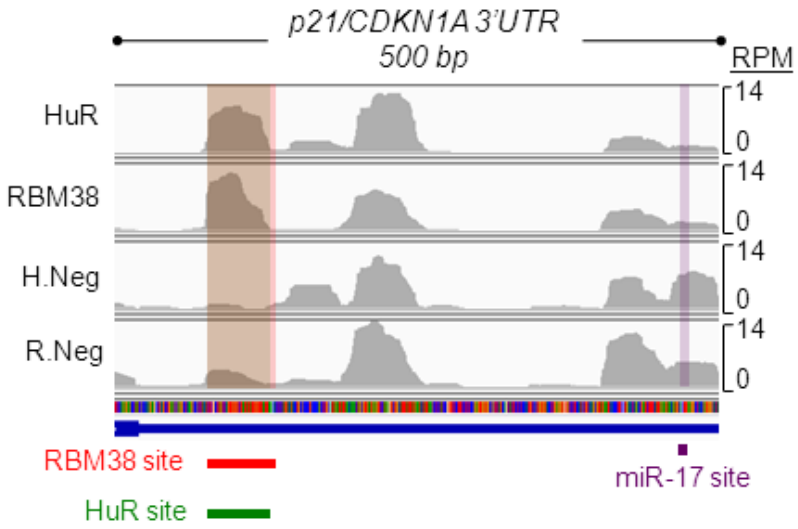

Figure 17: RBM38 and HuR binding sites (red and green bars and shading respectively) in *p21* mRNA 3'UTR. This location was previously reported to be cooperatively bound by RBM38 and HuR (Cho et al. 2010). Also note the predicted miR-17 site downstream. H.Neg, negative control from HuR DO-RIP-seq; R.Neg, negative control from RBM38 DO-RIP-seq.

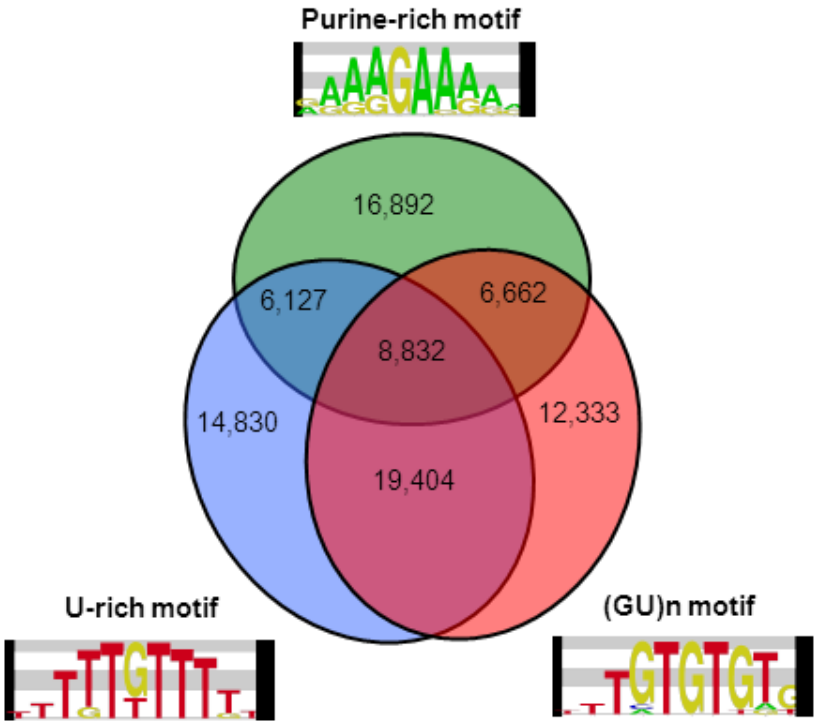

Figure 18: RBM38 binding sites have three RNA sequence motifs enriched. Note that overlapping regions of the Venn diagrams indicate binding sites containing two of the three, or all three RNA sequence motifs.

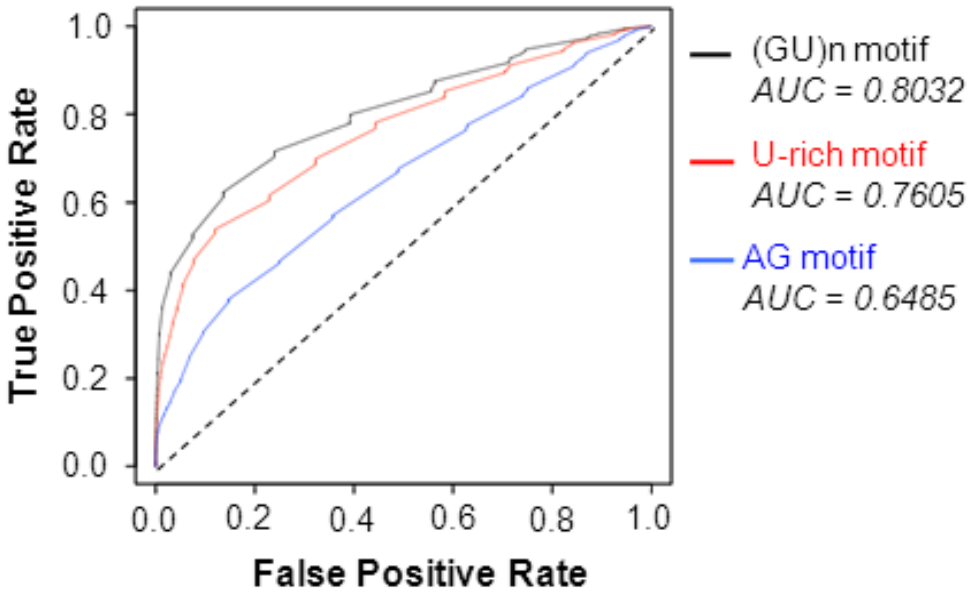

Figure 19: The presence of secondary structure affects RBM38 binding. AUROC analysis of predicting RBM38 binding using structure. AUROC scores suggest that RBM38 binding at (GU)n, U-rich, and AG/purine rich motifs is dependent on the lack of secondary structure at the location.

4.3.2 RBM38 binding sites frequently overlaps with HuR and TRA2B in a sequence-dependent manner

The GU-rich motif resembles sub-motifs detected in HuR binding sites from DO-RIP-seq **(Fig. 12A)**. Also, the locations of RBM38 binding sites in mRNA were similar to HuR **(Fig. 8B)**. So we hypothesized that RBM38 binding sites that contained GU-rich motifs were also bound by HuR in HEK293 cells. We bioinformatically defined sites bound by both RBM38 and HuR as overlapping DO-RIP-seq binding sites of each protein. Performing this analysis yielded 25,390 endogenous RBM38 binding sites that overlapped HuR binding sites. When RBM38 was over-expressed the number of binding

sites overlapping HuR increased to 55,740. These overlapping binding sites included the one previously reported in *p21/CDKN1A* mRNA 3'UTR **[94] (Fig. 17)**. The overlapping binding sites of RBM38 and HuR were located predominantly in 3'UTRs and introns; the binding sites in these locations combining for more than 82% of all overlapping sites **(Fig. 20A)**. In addition the sequence motif in RBM38 binding sites overlapping HuR binding sites were distinct from non-overlapping RBM38 binding sites. Using RBM38 binding sites in the 3'UTR as an example, non-overlapping sites in 3'UTRs had a [GU]n motif and overlapping sites had a GU-rich motif **(Fig. 20B)**.

We relied on a study that used NMR to show cooperative binding of RNA by RBM38 and RBFOX in nematodes to identify a candidate RBP for binding to purine rich motifs in RBM38 DO-RIP-seq sites. A protein alignment was performed using the basic local alignment search tool (BLAST, **[156]**) to identify human proteins with significant homology the amino acid sequence in nematode RBFOX that interacts with RBM38. This analysis yielded candidates from three RBP families: RBFOX, Hu/ELAVL, and TRA2 **(Table 3)**. Of these three candidates, TRA2 proteins have been reported to bind purine rich sequences in RNA through *in vitro* and transcriptome-wide studies **[115, 116, 118]**. Therefore we decided to perform DO-RIP-seq to identify the binding sites of TRA2 proteins in HEK293 cells.

DO-RIP-seq was performed on endogenous TRA2beta (TRA2B) in HEK293 cells, and yielded 125,775 binding sites that were predominantly in the CDS of mRNA **(Fig.**

21A). The sequence motif most frequently detected in TRA2B binding sites was a purine-rich motif **(Fig. 21B)**. Both the sequence motif and the location of TRA2B binding sites were in agreement with previous studies **[116, 118]**. The sequence motif recognized by TRA2B was similar to the purine-rich motif observed in some RBM38 binding sites. So, we were curious to know what sequence motifs were present in RBM38 binding sites overlapping and TRA2B sites, and how the locations in mRNA were distributed. There were 7,821 RBM38 binding sites overlapping TRA2B binding sites and included one in *ATP5C1* mRNA CDS **(Fig. 22)**. Greater than 83% of RBM38 binding sites overlapping TRA2B binding sites were in the CDS (53.6%) and 3'UTRs (30.3%) **(Fig. 23A)**. Using RBM38 binding sites in the CDS as a test case, we found that those overlapping TRA2B binding sites indeed had a purine-rich motif **(Fig. 23B)**. Yet again, RBM38 binding sites that did not overlap TRA2B sites had [GU]n motifs **(Fig. 23B)**.

Given the extent to which RBM38's and HuR's binding sites overlapped we were curious about how frequently all three RBPs (RBM38, HuR, and TRA2B) were detected binding to the same sites. Also, how did overlapping TRA2B and HuR binding sites compare to binding sites overlapping TRAB and RBM38 sites? To address these questions, we first compared the proportion of RBM38 binding sites overlapping HuR, TRA2B, or both HuR and TRA2B binding sites. These comparisons were performed on RBM38 binding sites classified by their locations in mRNA, i.e. sites located in 5'UTR, introns, CDS, or 3'UTR. We found in all mRNA locations that sites bound by all three

proteins were most infrequent **(Fig. 24)**. In addition we compared the proportion HuR- or TRA2B-bound sequence motifs overlapping RBM38 binding sites to non-bound motifs overlapping RBM38 sites. We defined bound sequence motifs as sites in mRNA containing HuR or TRA2B sequence motifs that were detected by DO-RIP-seq experiments of HuR and TRA2B respectively. We used motifs in CDS and 3'UTR locations. We found in both the CDS and in 3'UTRs that the proportion of RBM38 binding sites overlapping HuR-bound motifs or TRA2B-bound motifs were greater than the proportion overlapping non-bound motifs **(Fig. 25)**. We found a similar relationship for HuR binding sites **(Fig. 26)** and TRA2B binding sites **(Fig. 27)** overlapping RBM38-bound motifs when compared to non-bound RBM motifs. Interestingly, the HuR and TRA2B binding sites did not display this relationship; the difference between the proportions of HuR or TRA2B binding sites overlapping bound or non-bound TRA2B and HuR motifs respectively were much less in comparison to RBM38 motifs **(Fig. 26 and 27)**.

Altogether the integrative analysis of RBM38, HuR, and TRA2B transcriptome-wide binding revealed extensive co-binding that is RBM38-centric with HuR or TRA2B. By comparison, co-binding of sites in mRNA for HuR and TRA2B were much less extensive. Also apparent is the sequence dependence for co-binding of sites in mRNA by these proteins; non-overlapping binding sites of RBM38 had distinct sequence motifs from overlapping binding sites.

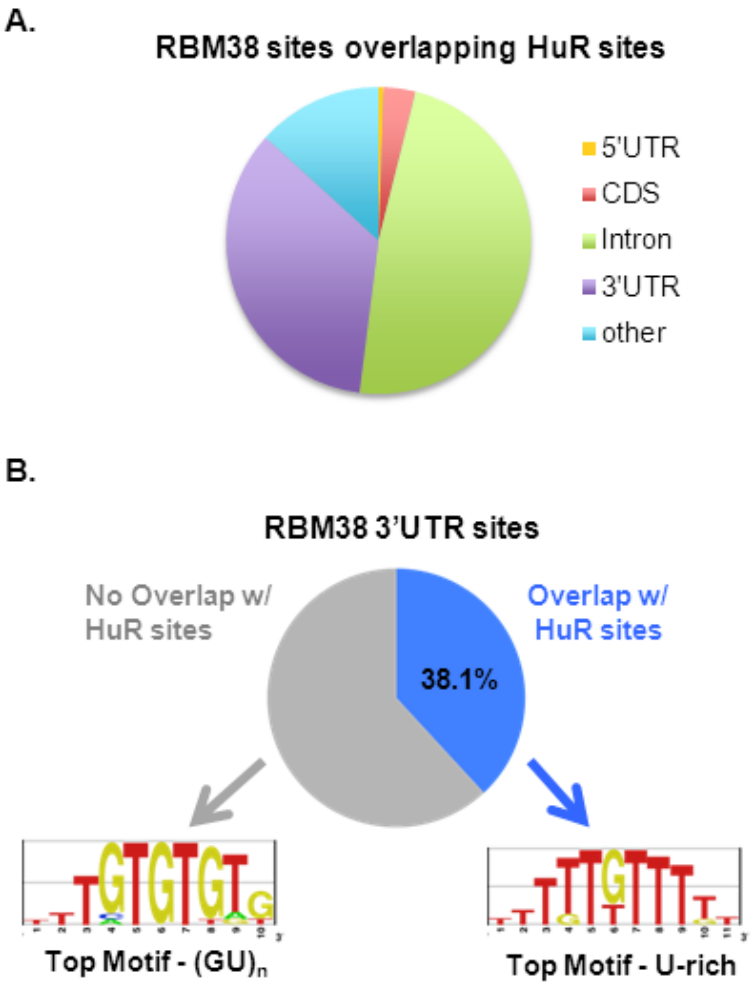

Figure 20: There is extensive evidence of overlapping binding sites for RBM38 and HuR in a sequence-specific manner. (A) Overlapping RBM38 and HuR binding sites were predominantly in 3'UTRs and introns of mRNA. (B). RBM38 sites overlapping HuR sites had a distinct GU-rich motif compared to the (GU)n motif observed in non-overlapping RBM38 sites.

Table 3: Protein BLAST identifies 3 human proteins with significant homology to the region of nematode Asd-1 (RBFOX) protein that interacts with nematode Sup-12 (RBM38).

Name	Max. Score	Coverage	E-value	Identity
RBFOX1	66.8	100%	4E-13	71%
HuR	36.3	88%	0.007	51%
TRA2	34.6	88%	0.026	47%

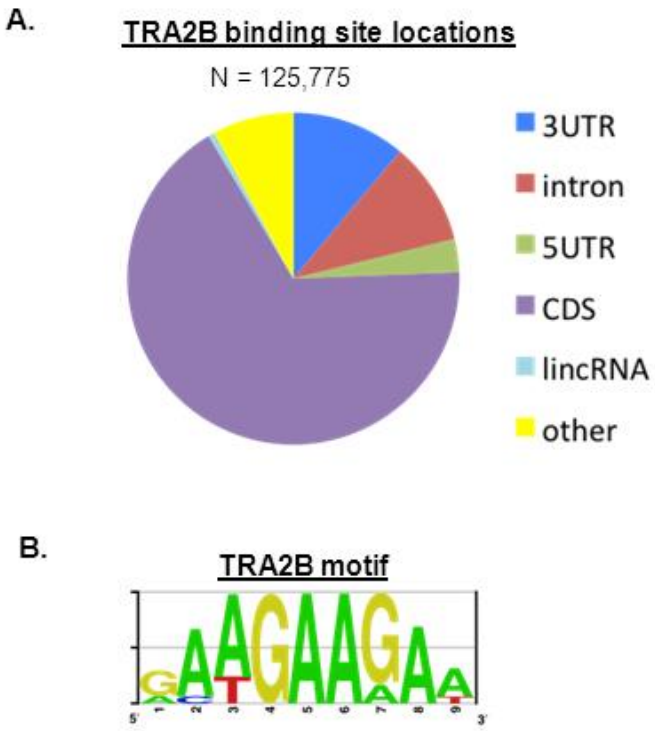

Figure 21: TRA2B DO-RIP-seq. (A) TRA2B binding sites were predominantly located in coding sequence exons (CDS, purple). (B) A purine-rich/AG-rich motif was most frequently detected in TRA2B binding sites.

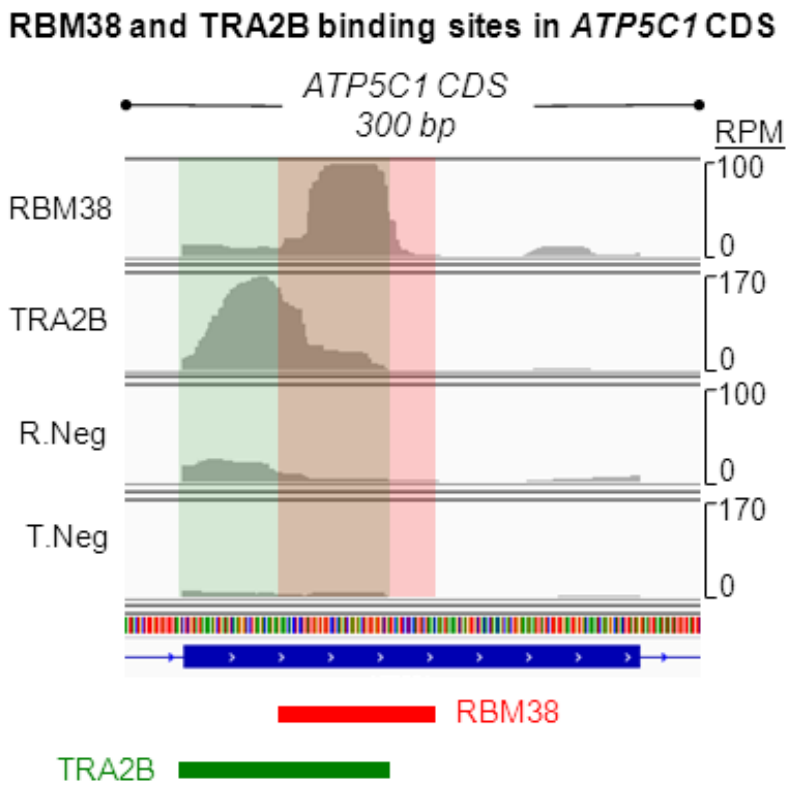

Figure 22: Overlapping binding sites of TRA2B (green) and RBM38 (red) in a CDS of the *ATP5C1* mRNA. R.Neg, negative control from RBM38 DO-RIP-seq; T.Neg, negative control from TRA2B DO-RIP-seq.

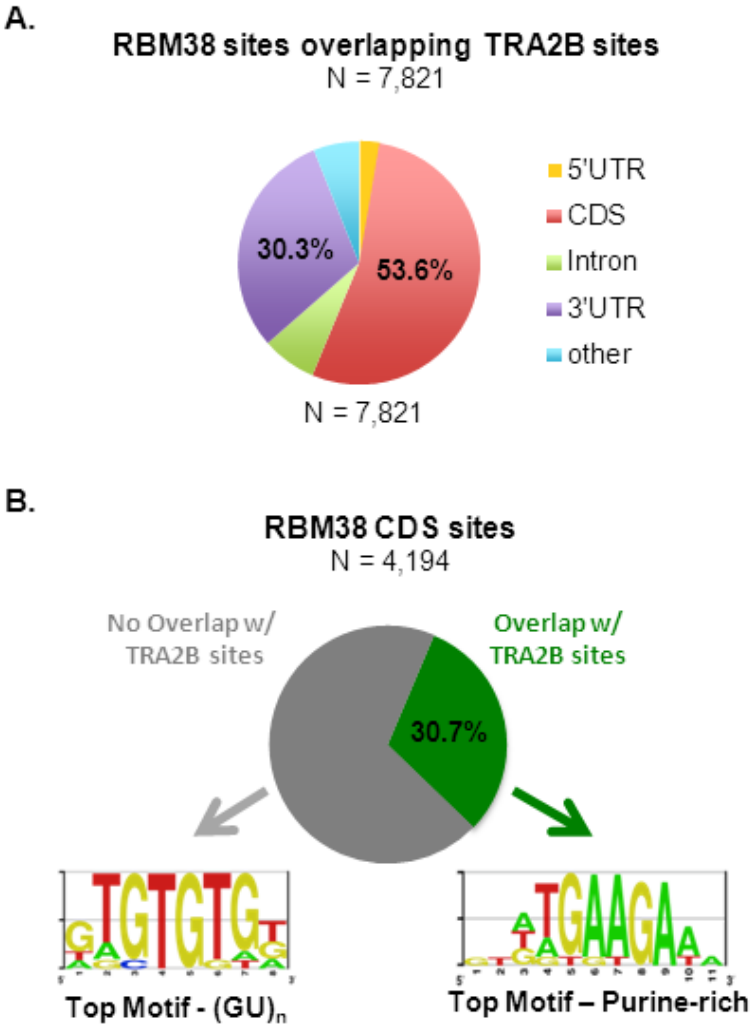

Figure 23: There is extensive evidence of overlapping binding sites for RBM38 and TRA2B in a sequence-specific manner. (A) Overlapping RBM38 and TRA2B binding sites were predominantly in CDS and 3'UTRs of mRNA. (B) RBM38 sites overlapping TRA2B sites had a distinct purine-/AG-rich motif compared to the (GU)n motif observed in non-overlapping RBM38 sites.

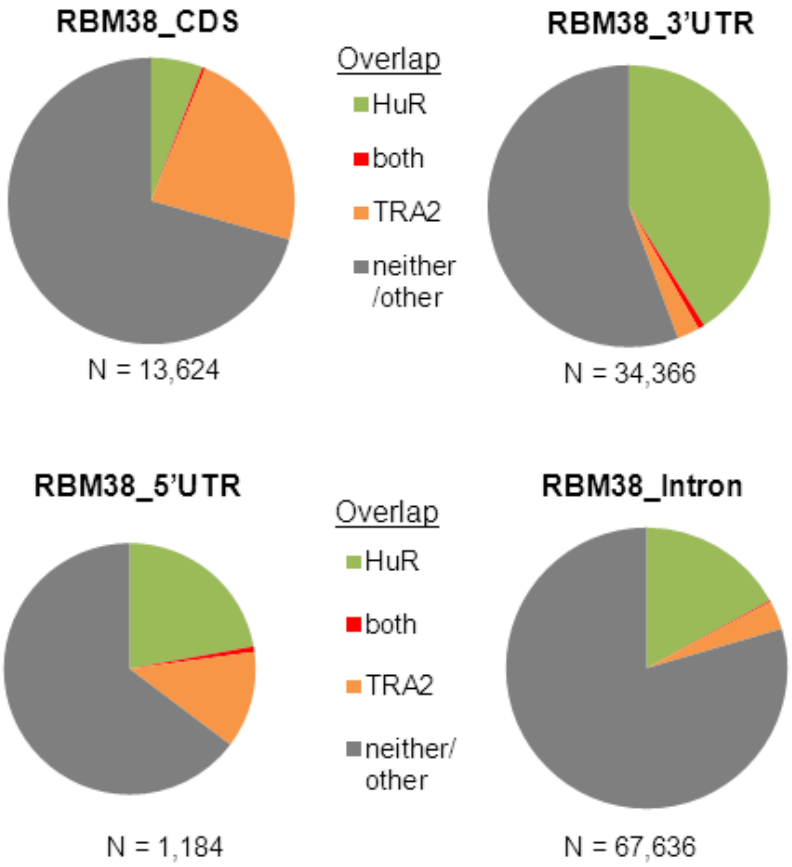

Figure 24: RBM38 binding sites overlap with HuR or TRA2B binding sites in all mRNA locations (CDS, 3'UTR, 5'UTR, Intron). However, all three proteins overlapping at the same locations (both, red slices) is relatively infrequent.

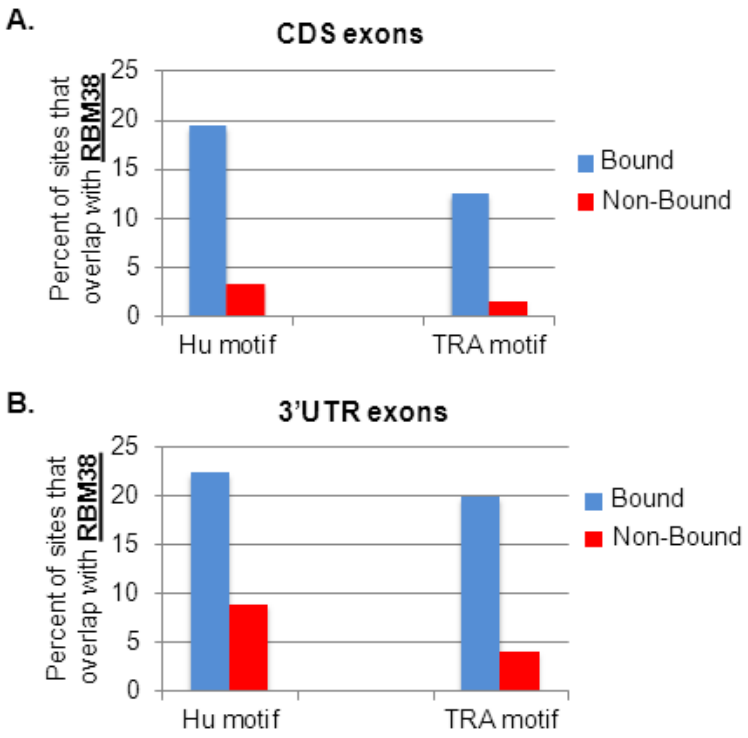

Figure 25: RBM38 binding sites more frequently overlap bound-HuR motifs and bound-TRA2B motifs than non-bound HuR or non-bound TRA2B motifs. This relationship is observed at (A) CDS and (B) 3'UTR sites in mRNA.

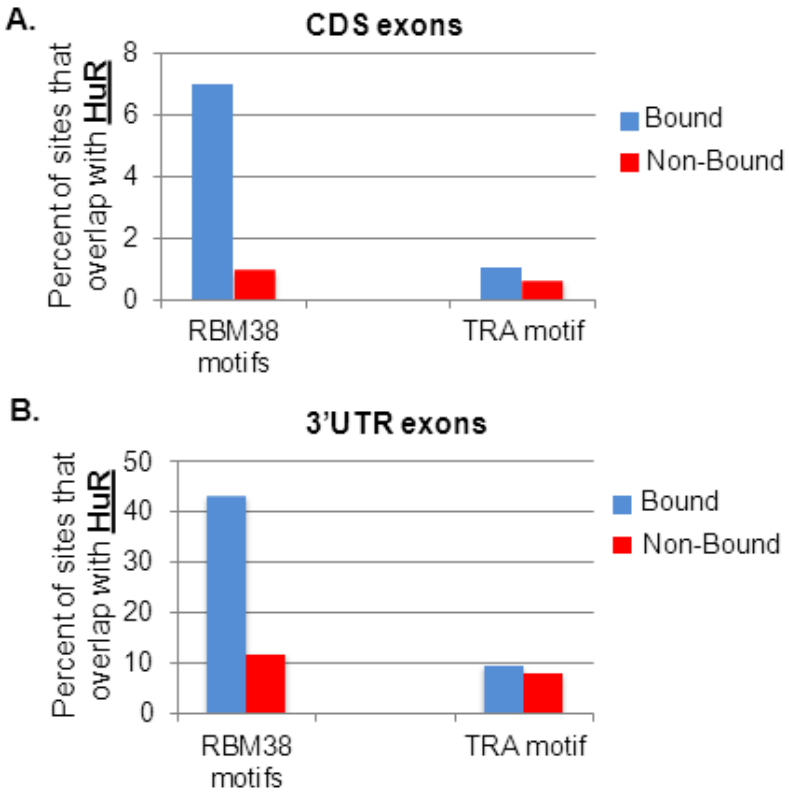

Figure 26: RBM38-bound motifs more frequently overlap HuR binding sites than non-bound RBM38 motifs. This is the case for sites in (A) CDS and (B) 3'UTRs. However the frequencies of TRA2-bound motifs overlapping HuR binding sites is not substantially greater than non-bound TRA2B motifs overlapping HuR sites in either (A) CDS or (B) 3'UTR.

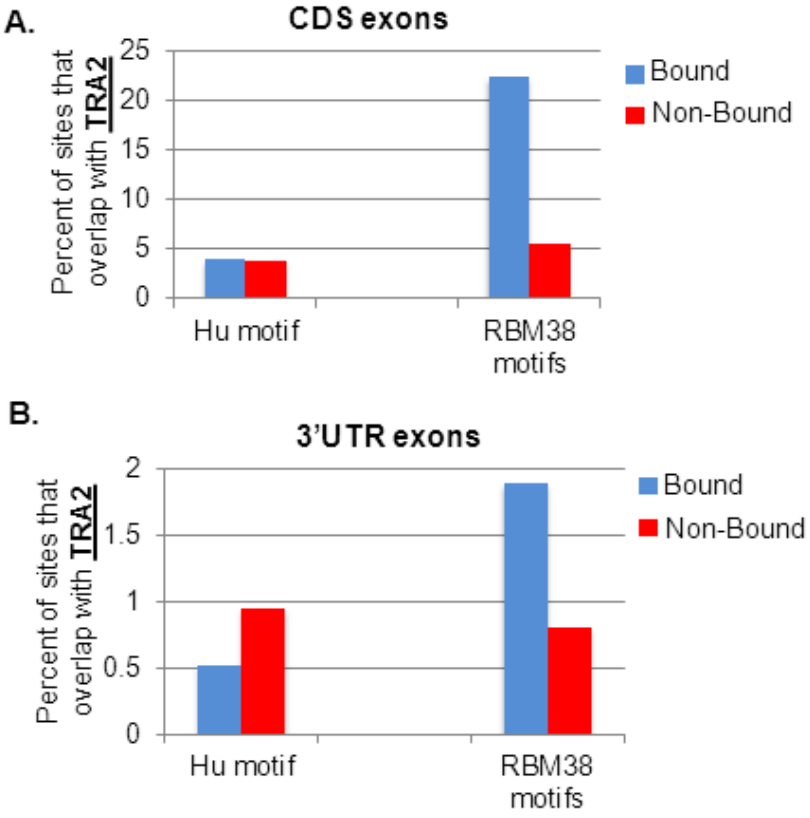

Figure 27: RBM38-bound motifs more frequently overlap TRA2B binding sites than RBM38 motifs that are non-bound. This is the case for sites in CDS (A) and 3'UTRs (B). However the frequencies of HuR-bound overlapping TRA2B binding sites is not greater than non-bound HuR motifs overlapping TRA2B sites in either CDS (A) or 3'UTR (B).

4.3.3 RBM38 binding sites that overlap HuR sites are the most conserved compared to non-overlapping and non-bound sites

Given the significant frequency of overlap between RBM38 and HuR binding sites, we wanted to know how the conservation between overlapping binding sites compared to non-overlapping binding sites and non-bound binding sites of RBM38. To do this we searched expressed 3'UTRs for sites containing any one of the three RBM38 sequence motifs ([GU]n, U-rich, or purine-rich). Next the average PhyloP scores were

calculated for motifs that were bound by RBM38, non-bound by RBM38, RBM38-bound and overlapping HuR sites, or RBM38-bound but not overlapping HuR sites. Positive PhyloP scores indicate that the nucleotide at genomic interval is under positive selection pressure, and negative PhyloP scores indicates that the nucleotide is under negative selection pressure **[157]**. We observed that RBM38-bound sites had more positive conservation scores than non-bound sites. In addition, RBM38-bound sites overlapping HuR sites had more positive conservation scores than RBM38-bound, non-overlapping sites **(Fig. 28)**. This trend was observed in the three RBM38 motif classes. Therefore, sites with RNA sequences containing RBM38 and HuR recognized sequence -elements were under selective pressure during evolution.

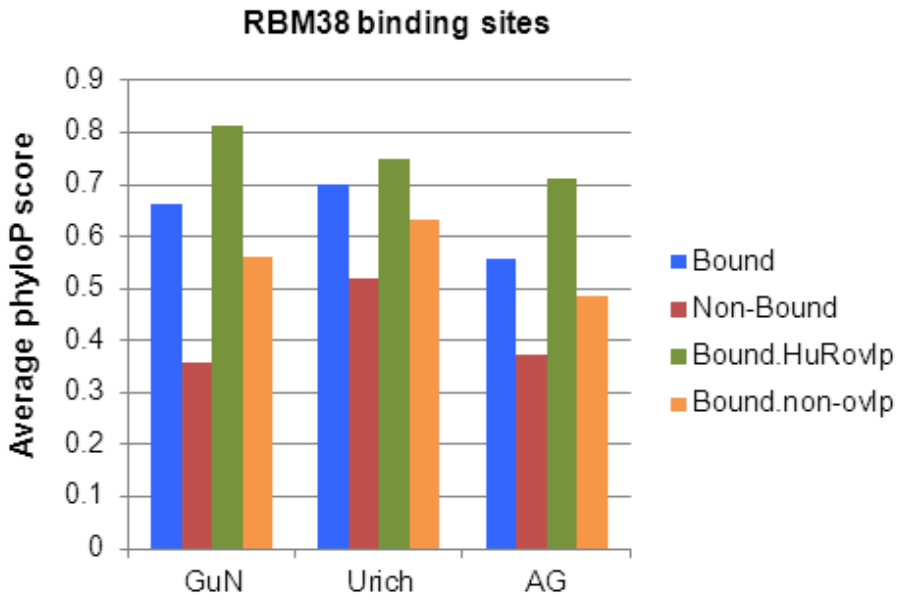

Figure 28: The average phyloP scores for nucleotides in sites containing (GU)n, U-rich, or AG-/purine-rich motifs are displayed. These sites were further categorized

into bound by RBM38 (blue), not bound by RBM38 (red), RBM38-bound and overlapping HuR sites (green), RBM38-bound but not overlapping HuR.

4.3.4 HuR and TRA2B binding sites that overlap RBM38 sites have significantly higher enrichment scores

Given that RBM38 binding sites overlapped HuR and TRA2B binding sites with significant frequencies, we wanted to know if there were differences between the relative strength of binding at overlapping binding sites and non-overlapping binding sites. Were these sites evidence of widespread combinatorial interactions between RBM38 and these RBP's? To answer this question we compared the binding site scores between overlapping and non-overlapping binding sites. In particular we compared the cumulative distribution functions of HuR or TRA2B sites overlapping RBM38 sites to non-overlapping sites. This analysis yielded significantly greater LOD scores between HuR binding sites overlapping RBM38 **(Fig. 29A)**, and significantly greater fold enrichment scores for TRA2B binding sites overlapping RBM38 compared to non-overlapping binding sites **(Fig. 29B)**. This suggests the potential for stronger binding at HuR or TRA2B binding sites that RBM38 also binds.

Next, we considered whether there were protein-protein interactions that influenced the relatively stronger binding at sites where HuR or TRA2B overlapped with RBM38. HEK293 cells were transiently transfected with an expression vector encoding RBM38 with a triple FLAG epitope tag and then immunoprecipitations of RBM38 with anti-FLAG antibodies followed by western blots were performed to detect co-immunoprecipitation of HuR and TRA2B. Also lysates were treated with *Micrococcal*

nuclease to digest protein-bound RNA prior to immunoprecipitations. HuR was detected in the immunoprecipitation of RBM38, but TRA2B was not detected **(Fig. 30A)**. Next we sought to validate the co-immunoprecipitation of HUR and RBM38 by performing the experiment where HuR is immunoprecipitated with mouse monoclonal antibody, and then western blot performed to detect RBM38. However, RBM38 was not detected in the immunoprecipitation of HuR **(Fig. 30B)**. This suggests that HuR and RBM38 might interact with each other when binding to RNA. However, a validation of the interaction is required and thus the evidence of an interaction between RBM38 and HuR remains inconclusive.

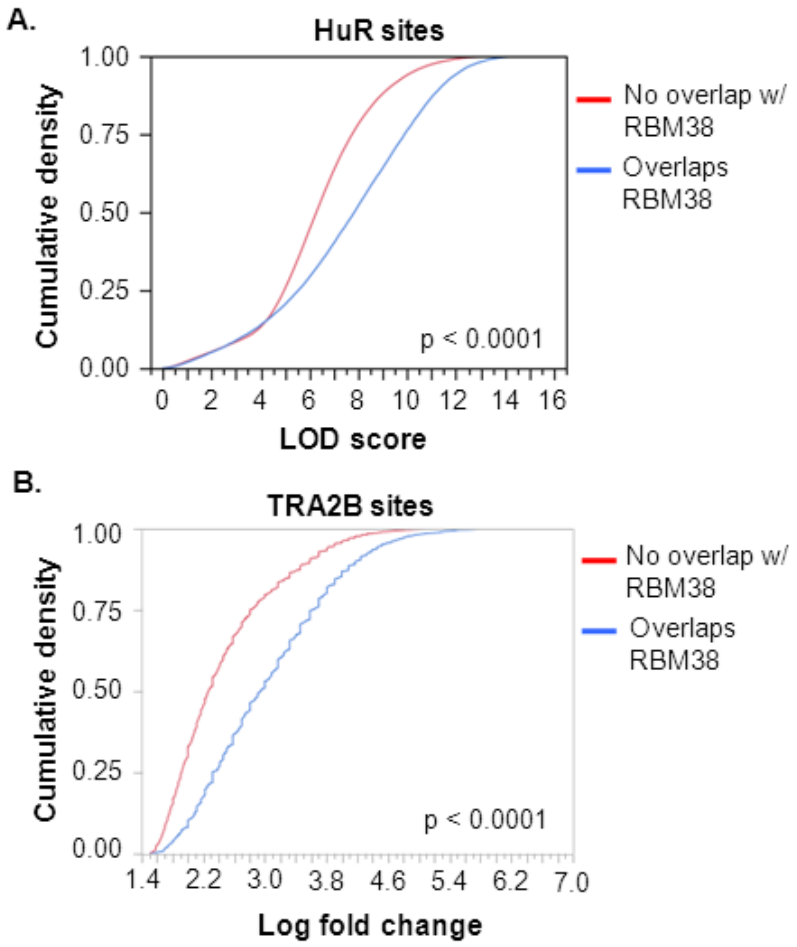

Figure 29: Cumulative distribution functions for (A) HuR sites and (B) TRA2B sites that overlapped or did not overlap RBM38 sites. HuR and TRA2B binding sites overlapping RBM38 sites had significantly higher LOD scores and log fold changes in enrichment respectively, than non-overlapping sites.

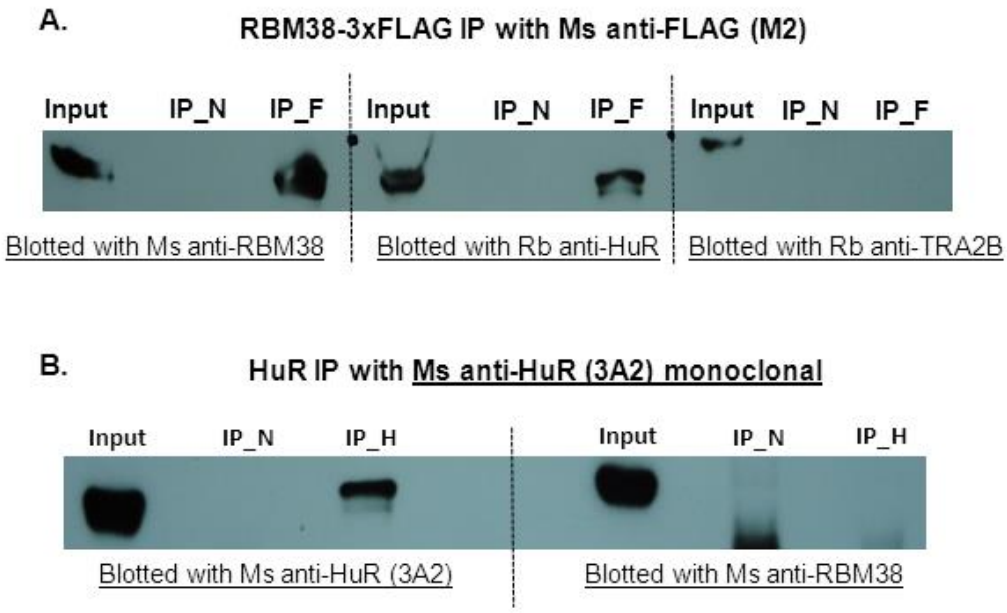

Figure 30: Immunoprecipitations (IPs) of (A) RBM38-3xFLAG or (B) HuR followed by western blots to detect the co-immunoprecipitations of HuR or TRA2B (with RBM38-3xFLAG IPs only). HuR was detected in RBM38 IPs while TRA2B was not. Conversely RBM38 was not detected in the IP of HuR.

4.3.5 Extensive overlap between the targets of RBM38 and HuR

Given the extensive transcriptome-wide overlap between RBM38 and HuR binding sites we were curious about the overlap between targets at whole transcript level. So, we performed an RSL analysis to generate RSL scores for transcript targets of RBM38. Next we defined HuR whole transcript targets as ones possessing an RSL score > 0 in the HuR RSL analysis. We found that HuR targets were shifted to the right in the RBM38 RSL score distribution **(Fig. 31A and 31B)**. Furthermore, we found that HuR targets had significantly greater RBM38 RSL scores than HuR non-targets (HuR RSLS score < 0) suggesting that RBM38 share similar targets. In fact, of the 48% of the

expressed mRNA that were targets of either RBM38 or HuR, two-thirds (26% of all expressed mRNA) were targets of both proteins **(Fig. 31C)**. Therefore, there is extensive overlap between the post-transcriptional targets of HuR and RBM38.

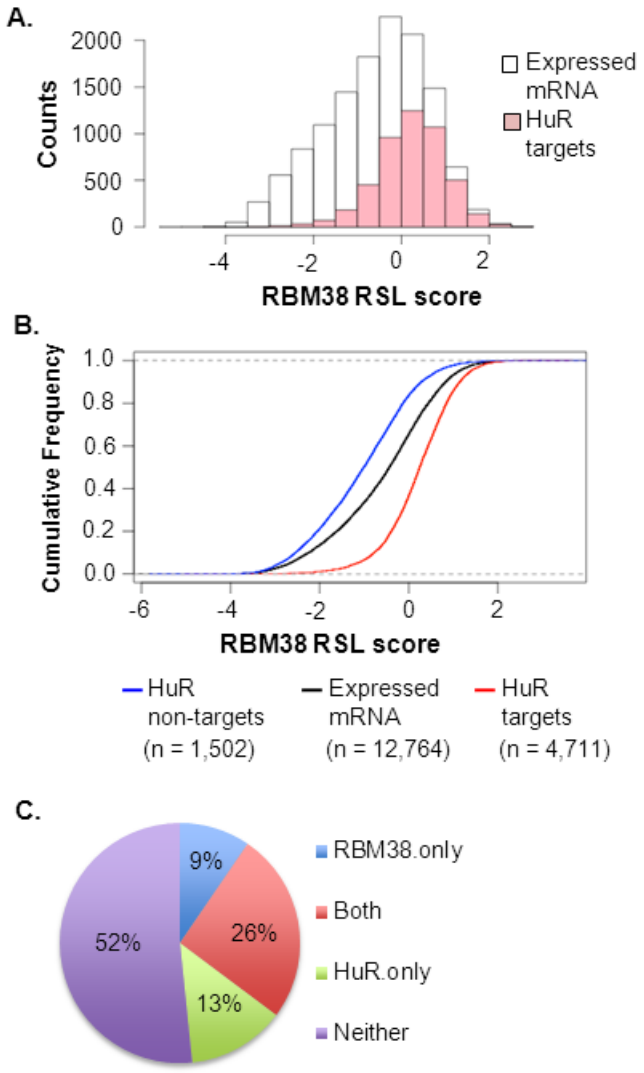

Figure 31: RBM38 associates more with HuR targets than HuR non-targets. (A) The distribution of RBM38 RSL scores. White bars are all expressed mRNA, and pink bars mRNAs that are HuR targets, i.e. HuR RSL score > 0. (B) Cumulative distribution function plots showing RBM38 RSL score distribution for HuR targets (HuR RSL score > 0, red), HuR non-targets (HuR RSL score < 0, blue), compared to all expressed

mRNA (black). (C) Venn diagram displaying comparing the proportions of expressed mRNAs that are targets of RBM38, HuR, or both RBPs.

4.3.6 Significantly enriched functional groups in RBM38 targets

Next we decided to look for the enrichment of functional groups in the whole-transcript targets of RBM38. Gene set enrichment analysis revealed that RBM38 targets were enriched for numerous miR targets **(Fig. 32)**. In fact, 96 miR target gene sets were significantly enriched with a FWER < 0.1. Given the substantial overlap between RBM38 and HuR targets we decided to measure the significance of enriched GO biological processes, and GO molecular functions in the co-targets of these RBPs. We found that some the most significantly enriched (FDR q-value < 0.05) biological processes were protein ubiquitination, WNT signaling and calcium modulating pathways, and the regulation of circadian rhythms among others **(Fig. 33)**. In terms of molecular functions, the most enriched were ubiquitin-like protein ligase activity, and core promoter sequence specific DNA binding **(Fig. 34)**. Interestingly, the only biological process significantly enriched in targets of RBM38 and not HuR was 'homophilic cell adhesion via plasma membrane adhesion molecules (data not shown). It must be noted that these mRNAs ('homophilic cell adhesion via plasma membrane adhesion molecules') are encoded by clustered protocadherin gamma genes. These genes encode exons from identical genomic loci and thus there appearance as an enriched group could be an artifact of how RSL scores are calculated. There were no biological process, or molecular functions significantly enriched in targets of HuR and not RBM38 (not shown).

Therefore, HuR and RBM38 potentially post-transcriptionally regulate similar biological processes in physiological conditions where both RBPs are expressed.

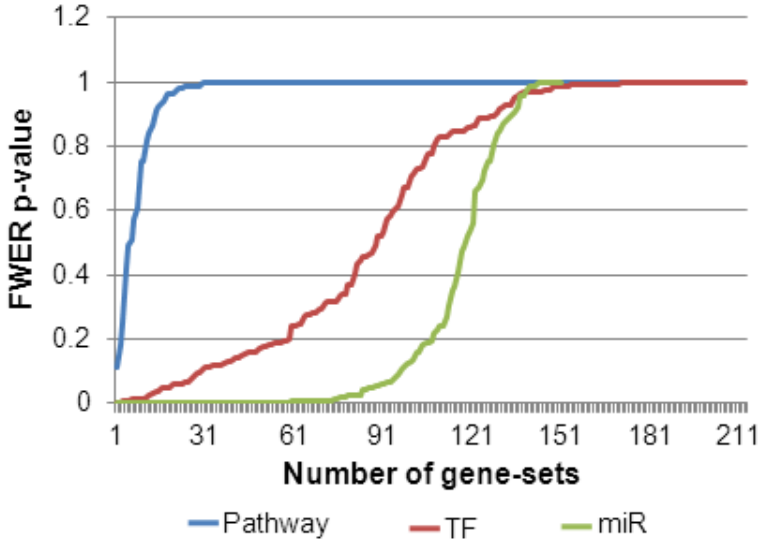

Figure 32: Genes encoding targets of microRNAs (miRs) are significantly enriched (FWER p-value < 0.1) in RBM38 whole transcript analysis according to gene set enrichment analysis (GSEA).

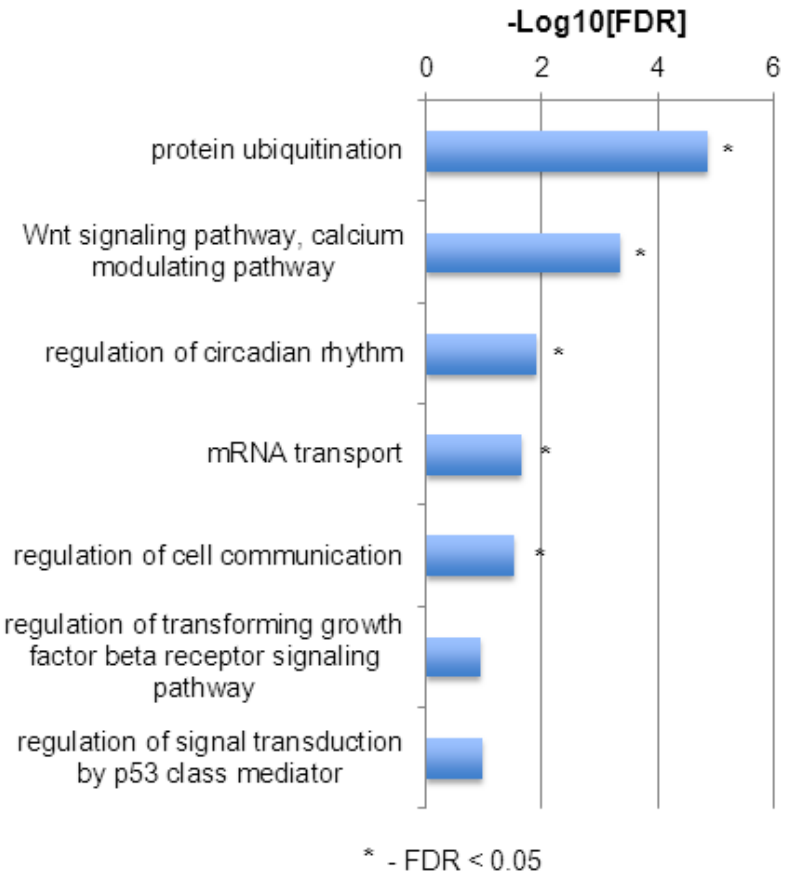

Figure 33: GO Biological Processes enriched in co-targets of RBM38 and HuR. The y-axis is the –log10 of the false discovery rate (FDR) q value for the calculated enrichments. An asterisk indicates a q value < 0.05.

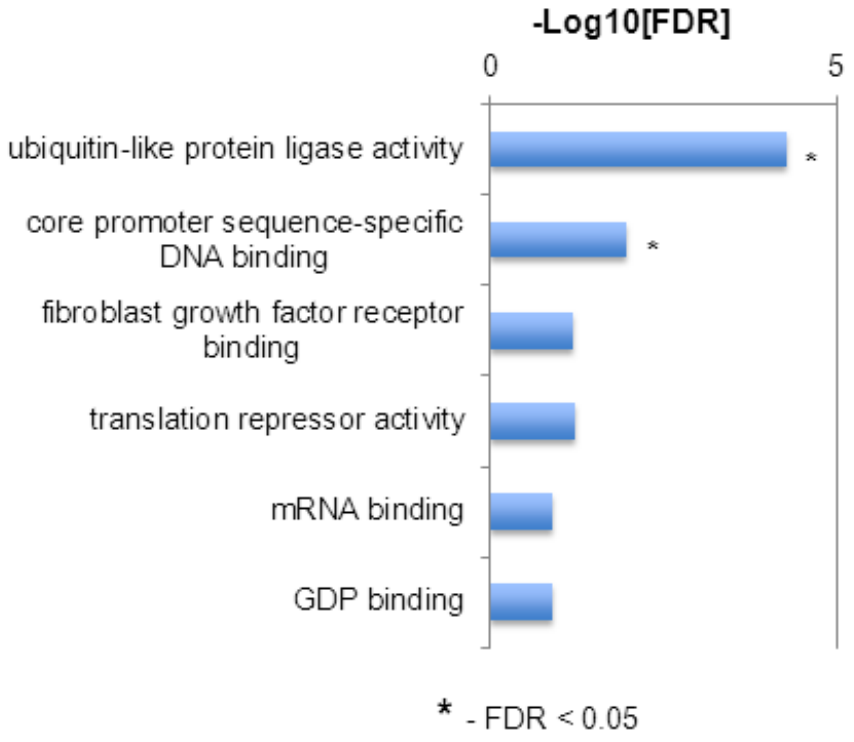

Figure 34: GO Molecular Functions enriched in co-targets of RBM38 and HuR. The y-axis is the –log10 of the false discovery rate (FDR) q value for the calculated enrichments. An asterisk indicates a q value < 0.05.

4.4 Discussion

The questions that drove the development of DO-RIP-seq were what were the sequence elements and protein partners that could regulate binding by an RBP? Our investigations of RBM38 binding with DO-RIP-seq revealed the significant co-occurrence of sequence motifs in RBM38 binding sites, and the requirement for single-stranded (non-structured) presentation of the motifs for binding. The recognition of sequence motifs in non-structured locations are attributes portrayed by most sequence specific RBPs **[77, 78, 158]** and so RBM38's binding of non-structured RNA motifs

strengthens the relationship between sequence-specificity and the binding of non-structured RNA. Also, given the extensive overlap between RBM38 and HuR, which is proven to bind non-structured RNA, RBM38's binding of non-structured RNA was expected.

The cooperative binding of mRNA by RBM38 has been reported by two independent studies. The first study to report this mechanism showed cooperation between the nematode homolog of RBM38 (SUP-12) and RBFOX (ASD-1) to bind an intronic site in egl-15 mRNA **[104]**. Another study reported the cooperative binding of a site in p21 mRNA by HuR and RBM38 **[94]**. Our study has greatly extended the number of transcriptomic locations where RBM38 could cooperate with HuR or TRA2B in a sequence specific manner. We observed RBM38 binding sites that contained RBFOX motifs but these occurred at lower frequencies than HuR and TRA2B motifs **(data not shown)**. Co-immunoprecipitation experiments provided some evidence for interactions between RBM38 and HuR, but there was none detected between RBM38 and TRA2B. It is possible that the most interactions between RBM38 and TRA2B occurs in the nucleus and thus is not readily detected in the mostly cytoplasmic cell lysate used. It is possible that nuclear lysates would yield different results for co-IPs of RBM38 and TRA2B. In addition RBM38 was not detected when HuR was immunoprecipitated, which contrasts with the IP of RBM38 **(Fig. 30)**. A monoclonal antibody was used to IP HuR for this experiment, and so it is possible that the epitope bound by the anti-HuR monoclonal

antibody is within or in proximity to the amino acids that interact with RBM38. The antibody could then disrupt the interaction between HuR and RBM38. To test this hypothesis, IPs of HuR would need to be performed with antibodies recognizing a different epitope followed by western blots to detect RBM38.

Another interesting finding was that bound RBM38 motifs were more conserved than non-bound ones. The higher conservation of bound compared to non-bound motifs coupled with the single-stranded requirement for RBM38 binding (and HuR binding) suggests an ancient mechanism of selection for RBP-recognized motifs amid flanking sequences with low probabilities of base-pairing. It also suggests that non-bound motifs are not involved in RBM38-mediate functions and so there is less fitness pressure to conserve the sequences in these locations. Though less conserved, it must be noted that the average phyloP scores for non-bound motifs were positive. Therefore, there is some selection for sequences flanking non-bound motifs with high base-pairing probabilities. The lower phyloP scores could result from changes in complementary nucleotides that would still maintain the secondary structure, and thus the motif remains non-bound.

In addition, bound RBM38 motifs overlapping HuR were more conserved than bound non-overlapping motifs **(Fig. 28)**. This suggests a greater positive selection for sequences that maintain the binding of RBM38 in close proximity to HuR in order to maintain the functional outcome of binding. As mentioned earlier, RBM38 is a tissue-specific regulator of alternative splicing and it can be induced in response to DNA

damage to stabilize mRNAs encoding cell cycle inhibitors **[153]**. It was demonstrated in mammalian cells that induction of RBM38 expression leads to increased stability of p21 mRNA due to cooperative binding with HuR **[94]**. It was also demonstrated in nematodes that tissue-specific expression of RBM38 leads to increased processing to yield a tissue specific egl-15 mRNA isoform through cooperative binding with RBFOX. The authors showed that disruption of the RBM38 or RBFOX motif abolished production of the tissue-specific egl-15 mRNA isoform. In addition the tissue-specific isoform produces a functionally distinct protein **[104]**. Therefore, these previous reports help explain the increased positive selection pressure on co-binding sites of RBM38 and HuR.

Our RSL score analysis of whole transcript targets of RBM38 is the first whole-transcript target analysis ever performed on RBM38. When compared to HuR whole transcript targets, it was revealed that the majority of HuR targets were also RBM38 targets. Given the large number of co-targets of these RBPs, it was no surprise that most RBM38 targets were also targets of numerous miRs. It was previously reported that RBM38's binding of mRNA can inhibit miR access **[95]**. Like HuR this might be due to the binding of RBM38 near the miR site, resulting in the direct competition for binding to the site. However, the RBM38 binding site in *p21* mRNA was hundreds of nucleotides away from the miR-17 binding sites. Inducing RBM38 levels was reported to reduce binding of AGO-miR-17 to *p21* mRNA **[95]**. Similarly, HuR's repression of stress induced miR-122 function on *SLC7A1* mRNA occurs at binding sites separated by

hundreds of nucleotides as well **[159]**. Therefore, the functional antagonism of miRs by RBM38 may not be limited to direct competition at the binding site.

To our knowledge, there has been no study revealing the mRNA targets of RBM38. RBM38 RSL analysis reveals that RBM38 shares a similar target set to HuR, and that these co-targets are significantly enriched in biological processes such as the regulation of circadian rhythms, WNT signaling and calcium modulation pathways, and protein ubiquitination. There have been previous reports on HuR's interaction with genes involved in circadian rhythms during T cell activation **[24]**, and *HuR* mRNA levels being regulated in a circadian manner in when macrophages are treated with lipopolysaccharide **[160]**. Liu and colleagues have reported a connection between HuR and the activation of WNT signaling pathway in promoting the growth of small intestinal mucosa in mice **[161]**. Liu and colleagues showed that mice lacking HuR had reduced protein levels of the WNT co-receptor LRP6, and that *LRP6* mRNA is a target of HuR. Also Mukherjee and colleagues reported the modulation of HuR binding to mRNAs encoding proteins involved in the WNT signaling pathway during a time-course of Jurkat T cell activation **[24]**. RBM38's targeting of mRNAs encoding proteins involved in the WNT signaling pathway has not been reported by other studies to our knowledge. WNT signaling pathway is involved in stem cell renewal as well as the differentiation of many cell types (reviewed in **[162-164]**) and thus RBM38 could post-transcriptionally regulate members of this pathway to facilitate these processes. HuR

and RBM38's targeting and potential regulation of protein ubiquitination pathway suggests that these RBPs may regulate protein turnover. Altogether, RBM38 target numerous processes that affect mRNA synthesis, processing, translation and protein turnover and thus could be considered a "regulator of regulators" (reviewed in **[165]**).

In conclusion RBM38 has the potential to modulate the binding of HuR and TRA2B in a sequence and secondary structure (the lack thereof) dependent manner. These binding relationships may be important for post-transcriptional regulation in conditions where RBM38 levels are increased in the presence of HuR and TRA2B such as erythropoiesis, myogenesis, and DNA damage. The transcriptome-wide binding sites and mRNA targets of RBM38 can be integrated with gene expression datasets to uncover functional targets of RBM38 involved in biological processes.

5. Concluding Remarks

The goal of my dissertation was to develop techniques that we could use to elucidate determinants of RBP binding transcriptome-wide. This goal was borne out of deficiencies in current techniques that lacked the ability to quantify binding of RNA at the whole transcript and binding site levels. These goals were met through the development of DO-RIP-seq. DO-RIP-seq melds aspects of RIP and CLIP techniques by using buffers to gently lyse the cells without disrupting RNP complexes, and through optimized RNA digestions with *Micrococcal* nuclease. I employed this strategy using HuR as a test case and obtained greatly improved transcriptomic coverage compared to CLIP, and reproducible binding sites. The increased transcriptome coverage allowed us to: 1) detect a greater number of functional HuR binding sites compared to CLIP techniques, 2) saturate binding site detection, and 3) distinguish features of bound versus non-bound HuR sites. Also, I achieved the ability to quantitatively detect binding; LOD scores for HuR binding sites correlated with relative binding strength and motif preference. The transcriptome-wide quantification of binding sites, to reveal preferential modes of binding was not achieved with previous techniques.

It is my belief that binding sites detection with DO-RIP-seq will allow for more granular comparisons between preferentially bound sites and regulation of the mRNAs stability or translation when integrated with transcriptome abundance and translation datasets. Obtaining more information on the determinants of RBP binding will improve

our ability to predict binding sites *in silico*. For example, a bioinformatics search for mRNAs containing the sequence motif recognized by an RBP would yield a large number of predicted sites; most of which are non-bound. However, by including base pairing probabilities for the nucleotides flanking the motif, one can better predict sites that are more likely to be bound by the RBP. Such an *in silico* approach would be useful when cellular material is limiting and/or antibodies against the RBP of interest are unavailable.

The ability to go beyond the detection of sequence motifs for recognition was not limited to HuR binding sites but also to other RBPs. The examination of RBM38 binding sites using DO-RIP-seq also revealed determinants of RBM38 binding and RBPs that frequently bind in proximity to RBM38; HuR and TRA2B in particular. The binding site scores at these "overlapping sites" revealed the potential for stronger binding and protein-protein interactions at these sites. Sites that allow the binding of multiple RBPs can have important functions, which is evident from previous reports and our analysis showing greater conservation (i.e. slower than expected rate of evolution) at RBM38 sites overlapping HuR. I believe that these qualities of DO-RIP-seq data will be advantageous for investigating changes in physiological states that modulate RBP binding in order to regulate gene expression post-transcriptionally. These qualities will also allow the discovery of RNA sequence determinants for ribonucleoprotein biogenesis involving multiple RBP partners. Roles for HuR, RBM38, and AGO-miRs in myogenesis have been

reported **[91, 104, 166-170]** and there is functional interplay between these factors. Given the extensive co-binding exhibited by HuR and RBM38, myogenesis is an intriguing process in which to study the modulation of HuR and AGO-miR binding when RBM38 is induced during myogenesis, and the effect on gene regulation.

Bibliography

1. Proudfoot, N.J., A. Furger, and M.J. Dye, *Integrating mRNA processing with transcription.* Cell, 2002. **108**(4): p. 501-12.

2. Bentley, D.L., *Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors.* Curr Opin Cell Biol, 2005. **17**(3): p. 251-6.

3. Kohler, A. and E. Hurt, *Exporting RNA from the nucleus to the cytoplasm.* Nat Rev Mol Cell Biol, 2007. **8**(10): p. 761-73.

4. Carmody, S.R. and S.R. Wente, *mRNA nuclear export at a glance.* J Cell Sci, 2009. **122**(Pt 12): p. 1933-7.

5. Gingras, A.C., B. Raught, and N. Sonenberg, *eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation.* Annu Rev Biochem, 1999. **68**: p. 913-63.

6. Jackson, R.J., C.U. Hellen, and T.V. Pestova, *The mechanism of eukaryotic translation initiation and principles of its regulation.* Nat Rev Mol Cell Biol, 2010. **11**(2): p. 113-27.

7. Wilusz, C.J., M. Wormington, and S.W. Peltz, *The cap-to-tail guide to mRNA turnover.* Nat Rev Mol Cell Biol, 2001. **2**(4): p. 237-46.

8. Sheth, U. and R. Parker, *Decapping and decay of messenger RNA occur in cytoplasmic processing bodies.* Science, 2003. **300**(5620): p. 805-8.

9. Schmid, M. and T.H. Jensen, *The exosome: a multipurpose RNA-decay machine.* Trends Biochem Sci, 2008. **33**(10): p. 501-10.

10. Keene, J.D., *Posttranscriptional generation of macromolecular complexes.* Mol Cell, 2003. **12**(6): p. 1347-9.

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

12. Keene, J.D., *Minireview: global regulation and dynamics of ribonucleic Acid.* Endocrinology, 2010. **151**(4): p. 1391-7.

13. Blackinton, J.G. and J.D. Keene, *Post-transcriptional RNA regulons affecting cell cycle and proliferation.* Semin Cell Dev Biol, 2014. **34C**: p. 44-54.

14. Jacob, F., et al., *[Operon: a group of genes with the expression coordinated by an operator].* C R Hebd Seances Acad Sci, 1960. **250**: p. 1727-9.

15. Futcher, B., et al., *A sampling of the yeast proteome.* Mol Cell Biol, 1999. **19**(11): p. 7357-68.

16. Gygi, S.P., et al., *Correlation between protein and mRNA abundance in yeast.* Mol Cell Biol, 1999. **19**(3): p. 1720-30.

17. Greenbaum, D., et al., *Comparing protein abundance and mRNA expression levels on a genomic scale.* Genome Biol, 2003. **4**(9): p. 117.

18. Tian, Q., et al., *Integrated genomic and proteomic analyses of gene expression in Mammalian cells.* Mol Cell Proteomics, 2004. **3**(10): p. 960-9.

19. Schmidt, M.W., et al., *Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe.* Mol Syst Biol, 2007. **3**: p. 79.

20. Tenenbaum, S.A., et al., *Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays.* Proc Natl Acad Sci U S A, 2000. **97**(26): p. 14085-90.

21. Schmitz-Linneweber, C., R. Williams-Carrier, and A. Barkan, *RNA immunoprecipitation and microarray analysis show a chloroplast Pentatricopeptide repeat protein to be associated with the 5′ region of mRNAs whose translation it activates.* Plant Cell, 2005. **17**(10): p. 2791-804.

22. Zhao, J., et al., *Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome.* Science, 2008. **322**(5902): p. 750-6.

23. Zhao, J., et al., *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Mol Cell, 2010. **40**(6): p. 939-53.

24. Mukherjee, N., et al., *Coordinated posttranscriptional mRNA population dynamics during T-cell activation.* Mol Syst Biol, 2009. **5**: p. 288.

25. Ule, J., et al., *CLIP identifies Nova-regulated RNA networks in the brain.* Science, 2003. **302**(5648): p. 1212-5.

26. Licatalosi, D.D., et al., *HITS-CLIP yields genome-wide insights into brain alternative RNA processing.* Nature, 2008. **456**(7221): p. 464-9.

27. Granneman, S., et al., *Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs.* Proc Natl Acad Sci U S A, 2009. **106**(24): p. 9613-8.

28. Granneman, S., et al., *Cracking pre-40S ribosomal subunit structure by systematic analyses of RNA-protein cross-linking.* EMBO J, 2010. **29**(12): p. 2026-36.

29. 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.

30. Hafner, M., et al., *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.* Cell, 2010. **141**(1): p. 129-41.

31. Konig, J., et al., *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.* Nat Struct Mol Biol, 2010. **17**(7): p. 909-15.

32. Ray, D., et al., *Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins.* Nat Biotechnol, 2009. **27**(7): p. 667-70.

33. Kishore, S., et al., *A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.* Nat Methods, 2011. **8**(7): p. 559-64.

34. Zarnack, K., et al., *Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements.* Cell, 2013. **152**(3): p. 453-66.

35. Friedersdorf, M.B. and J.D. Keene, *Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs.* Genome Biol, 2014. **15**(1): p. R2.

36. Jangi, M., et al., *Rbfox2 controls autoregulation in RNA-binding protein networks.* Genes Dev, 2014. **28**(6): p. 637-51.

37. Reyes-Herrera, P.H., et al., *BackCLIP: a tool to identify common background presence in PAR-CLIP datasets.* Bioinformatics, 2015. **31**(22): p. 3703-5.

38. Conway, A.E., et al., *Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival.* Cell Rep, 2016. **15**(3): p. 666-79.

39. Galloway, A., et al., *RNA-binding proteins ZFP36L1 and ZFP36L2 promote cell quiescence.* Science, 2016. **352**(6284): p. 453-9.

40. Holmqvist, E., et al., *Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo.* EMBO J, 2016. **35**(9): p. 991-1011.

41. Van Nostrand, E.L., et al., *Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP).* Nat Methods, 2016. **13**(6): p. 508-14.

42. Smith, K.C. and D.H. Meun, *Kinetics of the photochemical addition of [35S] cysteine to polynucleotides and nucleic acids.* Biochemistry, 1968. **7**(3): p. 1033-7.

43. Smith, K.C., *Photochemical addition of amino acids to 14C-uracil.* Biochem Biophys Res Commun, 1969. **34**(3): p. 354-7.

44. Sugimoto, Y., et al., *Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.* Genome Biol, 2012. **13**(8): p. R67.

45. Singh, G., E.P. Ricci, and M.J. Moore, *RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes.* Methods, 2014. **65**(3): p. 320-32.

46. Mili, S. and J.A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses.* RNA, 2004. **10**(11): p. 1692-4.

47. Keene, J.D., J.M. Komisarow, and M.B. Friedersdorf, *RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts.* Nat Protoc, 2006. **1**(1): p. 302-7.

48. Tenenbaum, S.A., et al., *Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays.* Methods, 2002. **26**(2): p. 191-8.

49. Penalva, L.O., et al., *RNA-binding proteins to assess gene expression states of co-cultivated cells in response to tumor cells.* Mol Cancer, 2004. **3**: p. 24.

50. van der Brug, M.P., et al., *RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways.* Proc Natl Acad Sci U S A, 2008. **105**(29): p. 10244-9.

51. Sanford, J.R., et al., *Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts.* Genome Res, 2009. **19**(3): p. 381-94.

52. Kershner, A.M. and J. Kimble, *Genome-wide analysis of mRNA targets for Caenorhabditis elegans FBF, a conserved stem cell regulator.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3936-41.

53. Jungkamp, A.C., et al., *In vivo and transcriptome-wide identification of RNA binding protein target sites.* Mol Cell, 2011. **44**(5): p. 828-40.

54. Mukherjee, N., et al., *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.

55. Wright, J.E., et al., *A quantitative RNA code for mRNA target selection by the germline fate determinant GLD-1.* EMBO J, 2011. **30**(3): p. 533-45.

56. Brooks, L., 3rd, et al., *A multiprotein occupancy map of the mRNP on the 3' end of histone mRNAs.* RNA, 2015. **21**(11): p. 1943-65.

57. Hansen, H.T., et al., *Drosophila Imp iCLIP identifies an RNA assemblage coordinating F-actin formation.* Genome Biol, 2015. **16**: p. 123.

58. Prasad, A., et al., *The PUF binding landscape in metazoan germ cells.* RNA, 2016.

59. Uren, P.J., et al., *High-throughput analyses of hnRNP H1 dissects its multi-functional aspect.* RNA Biol, 2016. **13**(4): p. 400-11.

60. Maris, C., C. Dominguez, and F.H. Allain, *The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.* FEBS J, 2005. **272**(9): p. 2118-31.

61. Bandziulis, R.J., M.S. Swanson, and G. Dreyfuss, *RNA-binding proteins as developmental regulators.* Genes Dev, 1989. **3**(4): p. 431-7.

62. Kenan, D.J., C.C. Query, and J.D. Keene, *RNA recognition: towards identifying determinants of specificity.* Trends Biochem Sci, 1991. **16**(6): p. 214-20.

63. Birney, E., S. Kumar, and A.R. Krainer, *Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors.* Nucleic Acids Res, 1993. **21**(25): p. 5803-16.

64. Auweter, S.D., F.C. Oberstrass, and F.H. Allain, *Sequence-specific binding of single-stranded RNA: is there a code for recognition?* Nucleic Acids Res, 2006. **34**(17): p. 4943-59.

65. Shamoo, Y., et al., *Crystal structure of the two RNA binding domains of human hnRNP A1 at 1.75 A resolution.* Nat Struct Biol, 1997. **4**(3): p. 215-22.

66. Xu, R.M., et al., *Crystal structure of human UP1, the domain of hnRNP A1 that contains two RNA-recognition motifs.* Structure, 1997. **5**(4): p. 559-70.

67. Deo, R.C., et al., *Recognition of polyadenylate RNA by the poly(A)-binding protein.* Cell, 1999. **98**(6): p. 835-45.

68. Handa, N., et al., *Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein.* Nature, 1999. **398**(6728): p. 579-85.

69. Allain, F.H., et al., *Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin.* EMBO J, 2000. **19**(24): p. 6870-81.

70. Wang, X. and T.M. Tanaka Hall, *Structural basis for recognition of AU-rich element RNA by the HuD protein.* Nat Struct Biol, 2001. **8**(2): p. 141-5.

71. van Gelder, C.W., et al., *A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation.* EMBO J, 1993. **12**(13): p. 5191-200.

72. Varani, L., et al., *The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein.* Nat Struct Biol, 2000. **7**(4): p. 329-35.

73. Levine, T.D., et al., *Hel-N1: an autoimmune RNA-binding protein with specificity for 3' uridylate-rich untranslated regions of growth factor mRNAs.* Mol Cell Biol, 1993. **13**(6): p. 3494-504.

74. Myer, V.E., X.C. Fan, and J.A. Steitz, *Identification of HuR as a protein implicated in AUUUA-mediated mRNA decay.* EMBO J, 1997. **16**(8): p. 2130-9.

75. Lopez de Silanes, I., et al., *Identification of a target RNA motif for RNA-binding protein HuR.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2987-92.

76. Meisner, N.C., et al., *mRNA openers and closers: modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure.* Chembiochem, 2004. **5**(10): p. 1432-47.

77. Kazan, H., et al., *RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins.* PLoS Comput Biol, 2010. **6**: p. e1000832.

78. Li, X., et al., *Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure.* RNA, 2010. **16**(6): p. 1096-107.

79. Gao, F.B., et al., *Selection of a subset of mRNAs from combinatorial 3' untranslated region libraries using neuronal RNA-binding protein Hel-N1.* Proc Natl Acad Sci U S A, 1994. **91**(23): p. 11207-11.

80. Jain, R.G., et al., *Hel-N1, an RNA-binding protein, is a ligand for an A + U rich region of the GLUT1 3' UTR.* Nucleic Acids Symp Ser, 1995(33): p. 209-11.

81. Jain, R.G., et al., *Ectopic expression of Hel-N1, an RNA-binding protein, increases glucose transporter (GLUT1) expression in 3T3-L1 adipocytes.* Mol Cell Biol, 1997. **17**(2): p. 954-62.

82. Fan, X.C. and J.A. Steitz, *Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs.* EMBO J, 1998. **17**(12): p. 3448-60.

83. Levy, N.S., et al., *Hypoxic stabilization of vascular endothelial growth factor mRNA by the RNA-binding protein HuR.* J Biol Chem, 1998. **273**(11): p. 6417-23.

84. Peng, S.S., et al., *RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein.* EMBO J, 1998. **17**(12): p. 3461-70.

85. Mazan-Mamczarz, K., et al., *RNA-binding protein HuR enhances p53 translation in response to ultraviolet light irradiation.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8354-9.

86. Bhattacharyya, S.N., et al., *Relief of microRNA-mediated translational repression in human cells subjected to stress.* Cell, 2006. **125**(6): p. 1111-24.

87. Lu, Y.C., et al., *ELAVL1 modulates transcriptome-wide miRNA binding in murine macrophages.* Cell Rep, 2014. **9**(6): p. 2330-43.

88. Kim, H.H., et al., *HuR recruits let-7/RISC to repress c-Myc expression.* Genes Dev, 2009. **23**(15): p. 1743-8.

89. Chen, X., et al., *Isolation and characterization of fourteen novel putative and nine known target genes of the p53 family.* Cancer Biol Ther, 2003. **2**(1): p. 55-62.

90. Heinicke, L.A., et al., *The RNA binding protein RBM38 (RNPC1) regulates splicing during late erythroid differentiation.* PLoS One, 2013. **8**(10): p. e78031.

91. Miyamoto, S., et al., *RNA-binding proteins Rbm38 and Rbm24 regulate myogenic differentiation via p21-dependent and -independent regulatory pathways.* Genes Cells, 2009. **14**(11): p. 1241-52.

92. Zhang, J., et al., *Mice deficient in Rbm38, a target of the p53 family, are susceptible to accelerated aging and spontaneous tumors.* Proc Natl Acad Sci U S A, 2014. **111**(52): p. 18637-42.

93. Ray, D., et al., *A compendium of RNA-binding motifs for decoding gene regulation.* Nature, 2013. **499**(7457): p. 172-7.

94. Cho, S.J., J. Zhang, and X. Chen, *RNPC1 modulates the RNA-binding activity of, and cooperates with, HuR to regulate p21 mRNA stability.* Nucleic Acids Res, 2010. **38**(7): p. 2256-67.

95. Leveille, N., et al., *Selective inhibition of microRNA accessibility by RBM38 is required for p53 activity.* Nat Commun, 2011. **2**: p. 513.

96. Zhang, J., S. Jun Cho, and X. Chen, *RNPC1, an RNA-binding protein and a target of the p53 family, regulates p63 expression through mRNA stability.* Proc Natl Acad Sci U S A, 2010. **107**(21): p. 9614-9.

97. Zhang, J., et al., *Translational repression of p53 by RNPC1, a p53 target overexpressed in lymphomas.* Genes Dev, 2011. **25**(14): p. 1528-43.

98. Cho, S.J., et al., *The RNA-binding protein RNPC1 stabilizes the mRNA encoding the RNA-binding protein HuR and cooperates with HuR to suppress cell proliferation.* J Biol Chem, 2012. **287**(18): p. 14535-44.

99. Xu, E., J. Zhang, and X. Chen, *MDM2 expression is repressed by the RNA-binding protein RNPC1 via mRNA stability.* Oncogene, 2013. **32**(17): p. 2169-78.

100. Yin, T., S.J. Cho, and X. Chen, *RNPC1, an RNA-binding protein and a p53 target, regulates macrophage inhibitory cytokine-1 (MIC-1) expression through mRNA stability.* J Biol Chem, 2013. **288**(33): p. 23680-6.

101. Zhang, J., E. Xu, and X. Chen, *Regulation of Mdm2 mRNA stability by RNA-binding protein RNPC1.* Oncotarget, 2013. **4**(8): p. 1121-2.

102. Cho, S.J., et al., *Hypoxia-inducible factor 1 alpha is regulated by RBM38, a RNA-binding protein and a p53 family target, via mRNA translation.* Oncotarget, 2015. **6**(1): p. 305-16.

103. Zhang, M., et al., *PPM1D phosphatase, a target of p53 and RBM38 RNA-binding protein, inhibits p53 mRNA translation via dephosphorylation of RBM38.* Oncogene, 2015.

104. Kuroyanagi, H., et al., *The Fox-1 family and SUP-12 coordinately regulate tissue-specific alternative splicing in vivo.* Mol Cell Biol, 2007. **27**(24): p. 8612-21.

105. Kuwasako, K., et al., *RBFOX and SUP-12 sandwich a G base to cooperatively regulate tissue-specific splicing.* Nat Struct Mol Biol, 2014. **21**(9): p. 778-86.

106. Belote, J.M. and B.S. Baker, *The dual functions of a sex determination gene in Drosophila melanogaster.* Dev Biol, 1983. **95**(2): p. 512-7.

107. Baker, B.S., *Sex in flies: the splice of life.* Nature, 1989. **340**(6234): p. 521-4.

108. Baker, B.S., et al., *Molecular genetic aspects of sex determination in Drosophila melanogaster.* Genome, 1989. **31**(2): p. 638-45.

109. Dauwalder, B., F. Amaya-Manzanares, and W. Mattox, *A human homologue of the Drosophila sex determination factor transformer-2 has conserved splicing regulatory functions.* Proc Natl Acad Sci U S A, 1996. **93**(17): p. 9004-9.

110. Beil, B., G. Screaton, and S. Stamm, *Molecular cloning of htra2-beta-1 and htra2-beta-2, two human homologs of tra-2 generated by alternative splicing.* DNA Cell Biol, 1997. **16**(6): p. 679-90.

111. Best, A., et al., *Expression of Tra2 beta in Cancer Cells as a Potential Contributory Factor to Neoplasia and Metastasis.* Int J Cell Biol, 2013. **2013**: p. 843781.

112. Hofmann, Y., et al., *Htra2-beta 1 stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2).* Proc Natl Acad Sci U S A, 2000. **97**(17): p. 9618-23.

113. Glatz, D.C., et al., *The alternative splicing of tau exon 10 and its regulatory proteins CLK2 and TRA2-BETA1 changes in sporadic Alzheimer's disease.* J Neurochem, 2006. **96**(3): p. 635-44.

114. Jiang, Z., et al., *Mutations in tau gene exon 10 associated with FTDP-17 alter the activity of an exonic splicing enhancer to interact with Tra2 beta.* J Biol Chem, 2003. **278**(21): p. 18997-9007.

115. Tacke, R., et al., *Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing.* Cell, 1998. **93**(1): p. 139-48.

116. Grellscheid, S., et al., *Identification of evolutionarily conserved exons as regulated targets for the splicing activator tra2beta in development.* PLoS Genet, 2011. **7**(12): p. e1002390.

117. Uren, P.J., et al., *Site identification in high-throughput RNA-protein interaction data.* Bioinformatics, 2012. **28**(23): p. 3013-20.

118. Best, A., et al., *Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons.* Nat Commun, 2014. **5**: p. 4760.

119. Tsuda, K., et al., *Structural basis for the dual RNA-recognition modes of human Tra2-beta RRM.* Nucleic Acids Res, 2011. **39**(4): p. 1538-53.

120. Ule, J., et al., *CLIP: a method for identifying protein-RNA interaction sites in living cells.* Methods, 2005. **37**(4): p. 376-86.

121. Heins, J.N., et al., *Characterization of a nuclease produced by Staphylococcus aureus.* J Biol Chem, 1967. **242**(5): p. 1016-20.

122. Noll, M. and R.D. Kornberg, *Action of micrococcal nuclease on chromatin and the location of histone H1.* J Mol Biol, 1977. **109**(3): p. 393-404.

123. Derrien, T., et al., *Fast computation and applications of genome mappability.* PLoS One, 2012. **7**(1): p. e30377.

124. Hafner, M., et al., *Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing.* Methods, 2008. **44**(1): p. 3-12.

125. Nicholson, C.O., M.B. Friedersdorf, and J.D. Keene, *Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq.* RNA, 2016.

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

127. Schueler, M., et al., *Differential protein occupancy profiling of the mRNA transcriptome.* Genome Biol, 2014. **15**(1): p. R15.

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

129. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

130. 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.

131. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.

132. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.

133. Smith, D.B. and L.M. Corcoran, *Expression and purification of glutathione-S-transferase fusion proteins.* Curr Protoc Mol Biol, 2001. **Chapter 16**: p. Unit16 7.

134. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.

135. Gao, F.B. and J.D. Keene, *Hel-N1/Hel-N2 proteins are bound to poly(A)+ mRNA in granular RNP structures and are implicated in neuronal differentiation.* J Cell Sci, 1996. **109 (Pt 3)**: p. 579-89.

136. Kullmann, M., et al., *ELAV/Hu proteins inhibit p27 translation via an IRES element in the p27 5'UTR.* Genes Dev, 2002. **16**(23): p. 3087-99.

137. Lal, A., et al., *Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs.* EMBO J, 2004. **23**(15): p. 3092-102.

138. Smola, M.J., et al., *SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells.* Proc Natl Acad Sci U S A, 2016. **113**(37): p. 10322-7.

139. Ma, W.J., et al., *Cloning and characterization of HuR, a ubiquitously expressed Elav-like protein.* J Biol Chem, 1996. **271**(14): p. 8144-51.

140. Lebedeva, S., et al., *Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR.* Mol Cell, 2011. **43**(3): p. 340-52.

141. Bernhart, S.H., I.L. Hofacker, and P.F. Stadler, *Local RNA base pairing probabilities in large sequences.* Bioinformatics, 2006. **22**(5): p. 614-5.

142. Dormoy-Raclet, V., et al., *The RNA-binding protein HuR promotes cell migration and cell invasion by stabilizing the beta-actin mRNA in a U-rich-element-dependent manner.* Mol Cell Biol, 2007. **27**(15): p. 5365-80.

143. Morris, A.R., N. Mukherjee, and J.D. Keene, *Ribonomic analysis of human Pum1 reveals cis-trans conservation across species despite evolution of diverse mRNA target sets.* Mol Cell Biol, 2008. **28**(12): p. 4093-103.

144. Kucukural, A., et al., *ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq.* Bioinformatics, 2013. **29**(19): p. 2485-6.

145. Lambert, N.J., A.D. Robertson, and C.B. Burge, *RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA-Binding Proteins.* Methods Enzymol, 2015. **558**: p. 465-93.

146. Campbell, Z.T. and M. Wickens, *Probing RNA-protein networks: biochemistry meets genomics.* Trends Biochem Sci, 2015. **40**(3): p. 157-64.

147. Lambert, N., et al., *RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins.* Mol Cell, 2014. **54**(5): p. 887-900.

148. Meisner, N.C. and W. Filipowicz, *Properties of the Regulatory RNA-Binding Protein HuR and its Role in Controlling miRNA Repression.* Adv Exp Med Biol, 2011. **700**: p. 106-23.

149. Srikantan, S., K. Tominaga, and M. Gorospe, *Functional interplay between RNA-binding protein HuR and microRNAs.* Curr Protein Pept Sci, 2012. **13**(4): p. 372-9.

150. Bushati, N. and S.M. Cohen, *microRNA functions.* Annu Rev Cell Dev Biol, 2007. **23**: p. 175-205.

151. Abdelmohsen, K. and M. Gorospe, *Posttranscriptional regulation of cancer traits by HuR.* Wiley Interdiscip Rev RNA, 2010. **1**(2): p. 214-29.

152. Stumpf, C.R., et al., *The translational landscape of the mammalian cell cycle.* Mol Cell, 2013. **52**(4): p. 574-82.

153. Shu, L., W. Yan, and X. Chen, *RNPC1, an RNA-binding protein and a target of the p53 family, is required for maintaining the stability of the basal and stress-induced p21 transcript.* Genes Dev, 2006. **20**(21): p. 2961-72.

154. Eden, E., et al., *Discovering motifs in ranked lists of DNA sequences.* PLoS Comput Biol, 2007. **3**(3): p. e39.

155. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC Bioinformatics, 2009. **10**: p. 48.

156. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

157. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.

158. Taliaferro, J.M., et al., *RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation.* Mol Cell, 2016. **64**(2): p. 294-306.

159. Kundu, P., et al., *HuR protein attenuates miRNA-mediated repression by promoting miRISC dissociation from the target RNA.* Nucleic Acids Res, 2012. **40**(11): p. 5088-100.

160. Keller, M., et al., *A circadian clock in macrophages controls inflammatory immune responses.* Proc Natl Acad Sci U S A, 2009. **106**(50): p. 21407-12.

161. Liu, L., et al., *RNA-binding protein HuR promotes growth of small intestinal mucosa by activating the Wnt signaling pathway.* Mol Biol Cell, 2014. **25**(21): p. 3308-18.

162. Logan, C.Y. and R. Nusse, *The Wnt signaling pathway in development and disease.* Annu Rev Cell Dev Biol, 2004. **20**: p. 781-810.

163. Clevers, H., *Wnt/beta-catenin signaling in development and disease.* Cell, 2006. **127**(3): p. 469-80.

164. Clevers, H., K.M. Loh, and R. Nusse, *Stem cell signaling. An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control.* Science, 2014. **346**(6205): p. 1248012.

165. Mansfield, K.D. and J.D. Keene, *The ribonome: a dominant force in co-ordinating gene expression.* Biol Cell, 2009. **101**(3): p. 169-81.

166. Figueroa, A., et al., *Role of HuR in skeletal myogenesis through coordinate regulation of muscle differentiation genes.* Mol Cell Biol, 2003. **23**(14): p. 4991-5004.

167. Deschenes-Furry, J., et al., *The RNA-binding protein HuR binds to acetylcholinesterase transcripts and regulates their expression in differentiating skeletal muscle cells.* J Biol Chem, 2005. **280**(27): p. 25361-8.

168. Chen, J.F., et al., *The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation.* Nat Genet, 2006. **38**(2): p. 228-33.

169. Rao, P.K., et al., *Myogenic factors that regulate expression of muscle-specific microRNAs.* Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8721-6.

170. Dormoy-Raclet, V., et al., *HuR and miR-1192 regulate myogenesis by modulating the translation of HMGB1 mRNA.* Nat Commun, 2013. **4**: p. 2388.

Biography

Cindo O. Nicholson was born in the parish of St. Andrew, in the Caribbean nation of Jamaica. Cindo obtained an Associates of Arts degree in General Studies from Edison College, and a Bachelor's of Science degree in Biotechnology from Florida Gulf Coast University, both located in Fort Myers, FL.

Publications at Duke:

Nicholson CO*, Friedersdorf MB*, Bisogno LS, Keene JD (2016). DO-RIP-seq to quantify RNA binding sites transcriptome-wide. METHODS, doi: 10.1016/j.ymeth.2016.11.004.

Nicholson CO*, Friedersdorf MB*, Keene JD (2016). Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. RNA, doi: 10.1261/rna.058115.116.

Smola MJ, Christy TW, Inoue K, Nicholson CO, Friedersdorf MB, Keene JD, Lee DM, Calabrese JM, Weeks KM (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proc natl Acad Sci USA 113(37): 10322-10327.

Tomaras GD, Ferrari G, Shen X, Alam SM, Liao HX et al. (2013). Vaccine-induced plasma IgA specific for the C1 region of the HIV-1 envelope blocks binding and effector function of IgG. Proc Natl Acad Sci USA 110(22): 9019-9024.