

2014

RNA Deregulation in Metastatic Breast Cancer

Christina Beate Marney

Follow this and additional works at: http://digitalcommons.rockefeller.edu/student_theses_and_dissertations

 Part of the [Life Sciences Commons](#)

Recommended Citation

Marney, Christina Beate, "RNA Deregulation in Metastatic Breast Cancer" (2014). *Student Theses and Dissertations*. Paper 262.

This Thesis is brought to you for free and open access by Digital Commons @ RU. It has been accepted for inclusion in Student Theses and Dissertations by an authorized administrator of Digital Commons @ RU. For more information, please contact mcsweej@mail.rockefeller.edu.



RNA DEREGLATION IN METASTATIC BREAST CANCER

A Thesis Presented to the Faculty of
The Rockefeller University
In Partial Fulfillment of the Requirements for
The degree of Doctor of Philosophy

By
Christina Beate Marney

June 2014

RNA DYSREGULATION IN METASTATIC BREAST CANCER

Christina Beate Marney, Ph.D.

The Rockefeller University 2014

Breast cancer is the most commonly diagnosed and second highest cause of cancer-related mortality in women, with the majority of deaths associated with metastatic spread of tumors to distal organs. The metastatic process involves multiple sequential steps including invasion of the surrounding stroma, intravasation into lymph or blood vessels, survival during transport with successful evasion of the immune system, extravasation and final outgrowth in the new microenvironment. At each step cells acquire new properties through the alteration of gene expression profiles. While a great deal of effort has been made towards understanding the underlying genetic mutations associated with enhanced motility and invasiveness, it is becoming increasingly apparent that post-transcriptional modification of gene expression at the RNA level contributes substantially towards gain of metastatic potential.

In eukaryotic cells RNA is subject to multiple layers of regulation that together generate extensive phenotypic and temporal complexity from a relatively small number of genes. RNA regulation is achieved through the concerted action of multiple RNA binding proteins (RBPs) which themselves are subject to tight regulation. Recent studies have implicated the aberrant expression of multiple RBPs in tumor

development and metastatic spread, while technological advances have provided methods by which unbiased, transcriptome-wide footprints of RBP-RNA interactions can be mapped. High-throughput cross-linked immunoprecipitation (HITS-CLIP) provides both the identity of regulated transcripts and the location of RBP regulation within a transcript. For the post-transcriptional regulatory protein Argonaute (AGO), which binds mRNA in complex with small non-coding microRNAs (miRNAs), HITS-CLIP derived 3'UTR binding sites have been shown to correspond to sites of functional miRNA-mediated target downregulation.

In this study, a molecularly characterized cell culture based model system of breast cancer metastasis has been employed to investigate the role of RNA binding proteins in the inhibition of metastatic progression. AGO-mRNA regulation mediated by endogenous miRNAs and exogenous anti-metastatic pre-miR-335 has been interrogated by HITS-CLIP to identify on a genome-wide scale differentially regulated targets with respect to metastatic potential. Furthermore, HITS-CLIP has been used to characterize a novel anti-metastatic RNA binding protein, RBM47, uncovering multifunctional regulation of target transcript stability and alternative splicing.

For the women behind the identifiers

ACKNOWLEDGEMENTS

I would like to start by thanking my advisor Dr. Robert B. Darnell for opening my eyes to the endless possibility in RNA biology. Thank you for your support, your humor and your willingness to always give the big picture.

I would like to thank the past a present members of the Darnell lab for each contributing towards making such an enjoyable work environment. From the emergency bottle of champagne in the cold room for “bad experiment days” to the in-depth discussions during lab meetings, your camaraderie, mentorship and willingness to help have been greatly appreciated. I would particularly like to thank Dr. Chaolin Zhang for his guidance and help with bioinformatics analyses, Aldo Mele for teaching me CLIP, and Jak Fak for being an RNA master.

I would like to thank the current and past members of my thesis committee, Drs. Allis, Tavazoie and Heintz for their advice and guidance over the years. I am also appreciative of Dr. Scott Lowe for participating as my outside committee member.

There are many people who have made my time at Rockefeller University so rewarding, this is a special place and I am proud to have been a member of this community. I would like to thank Dr. Charles Gilbert who first brought me into the Rockefeller fold as a technician, and all the

members of the Dean's Office for their support and guidance, in particular Dr. Emily Harms for her calm and reassuring demeanor. I couldn't have asked for a better group of people to call my classmates, I hope my Rockefeller friends know that they are all very dear to me.

I am lucky to have a diverse group of friends outside of the Rockefeller gates. I would like to thank my Sunday dinner chef Dr. Michael Raven, my knitters, my flickr-ers, my circus folk and my quilters for continually enriching my life outside of the lab.

Adam, I love you.

Finally, I would like to my mum and dad for not batting an eyelid when at 18 I moved across the country to UEA, and when at 23 I declared that I was moving to New York. Thank you for always supporting me, for giving me the freedom to explore and for always being my backup. To my big sister Gina and her family, Toll and William, thank you for giving me a reason to be homesick. It's a new and wonderful feeling.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS	vi
LIST OF FIGURES	xi
LIST OF TABLES	xvi
 CHAPTER 1: Introduction	 1
RNA Binding proteins and cancer-associated alterations in RNA metabolism	2
<i>Alternative Splicing</i>	3
<i>Cancer-associated mutations resulting alternative splicing</i>	5
<i>Splicing regulators as oncogenes – SRSF1</i>	6
<i>Splicing regulators as oncogenes – FOX2</i>	8
<i>Splicing regulators as tumor suppressors – RBM5</i>	9
<i>Alternative splicing in metastasis</i>	10
<i>RNA binding proteins and modulation of transcript stability</i>	12
<i>Nonsense mediated decay in cancer</i>	12
<i>Transcript stability, ARE-binding proteins and cancer</i>	15
<i>MicroRNA mediated control of gene expression in oncogenesis & metastasis</i>	17
<i>Alternative polyadenylation site selection</i>	20
Mapping transcriptome-wide regulatory networks of RNA binding proteins	22
Perspectives	25

CHAPTER 2: Experimental Procedures

<i>Cell Culture</i>	27
<i>Analysis of miRNA and mRNA expression levels by quantitative-PCR</i>	27
<i>AGO HITS-CLIP</i>	29
<i>RBM47 polyA-tailed HITS-CLIP</i>	36
<i>RNAseq Library Preparation</i>	41
<i>3'UTR Cloning into psiCheck2 and generation of mutants</i>	41
<i>MicroRNA mimic transfections & dual Luciferase 3'UTR Reporter Assays</i>	43
<i>Generation of CETN3 overexpressing stable cell lines</i>	43
<i>Invasion assays</i>	44
<i>Flow Cytometry Surface Staining</i>	45
<i>Western blotting</i>	45
<i>Alternative splicing semi-quantitative RT-PCR</i>	46

CHAPTER 3: Differential AGO-regulation in weakly and highly metastatic breast cancer.

Introduction	48
Results	55
<i>AGO-HITS CLIP in MDA-MB-231 and its highly aggressive metastatic derivatives with tissue tropism to the lung, bone and brain</i>	55
<i>AGO-miRNA profiles of MDA-231 and its metastatic sub-lines are highly similar</i>	61
<i>Identification of differentially CLIPed miRNA with respect to tissue tropism</i>	63
<i>Global analysis of AGO-mRNA binding</i>	65

<i>AGO-mRNA binding patterns are highly reproducible between cell lines</i>	69
<i>Bioinformatic identification of differential AGO binding sites</i>	71
<i>Identification of AGO-mRNA binding sites gained in metastatic derivatives of MDA-231</i>	77
<i>Identification of AGO-binding sites commonly lost in highly metastatic sub-lines</i>	79
<i>Differential AGO binding in the 3'UTR of HMGA1</i>	79
<i>Differential AGO binding in the 3'UTR of CD59</i>	88
Discussion	93

CHAPTER 4: Transcriptome-wide target identification of the anti-metastatic miR-335

Introduction	107
Results	111
<i>AGO HITS-CLIP in 4175LM2 Control and 4175LM2 pre-miR335 over-expressing cells</i>	111
<i>Characterization of the AGO-miRNA profiles of LM2 control and LM2 335OE</i>	111
<i>AGO HITS-CLIP reveals strand bias towards miR-335*</i>	114
<i>Global characterization of AGO-mRNA binding in LM2 Control and LM2 335OE</i>	121
<i>Identification of miR-335/* dependent de novo AGO-mRNA clusters</i>	124
<i>Identification of miR-335/* dependent de novo AGO-mRNA significant peaks</i>	131
<i>Initial selection of miR-335/* targets for validation</i>	134

<i>Validation of centrin EF-hand protein 3 (CETN3) as a miR-335* target</i>	137
<i>Clinical association of CETN3 and metastasis free survival</i>	143
<i>Further validation of predicted miR-335/* targets identified by AGO-HITS CLIP</i>	146
<i>Comparison of AGO HITS-CLIP derived de novo clusters miR-335/* targets and previously published targets</i>	150
Discussion	157

CHAPTER 5: Characterization of RBM47, an RNA-binding suppressor of breast cancer metastasis

Introduction	166
Results	169
<i>RBM47 expression is decreased in the metastatic derivatives of MDA-231</i>	169
<i>Wild-type RBM47 functions as a metastasis suppressor</i>	169
<i>Domain architecture of RBM47</i>	171
<i>Protein BLAST of RBM47 reveals sequence similarities to the known RNA-binding proteins</i>	175
<i>RBM47 is an RNA binding protein</i>	175
<i>Mapping direct RBM47-RNA interactions using a modified HITS-CLIP protocol</i>	179
<i>Summary of RBM47 HITS-CLIP data</i>	183
<i>Reproducibility of RBM47 HITS-CLIP data sets</i>	185
<i>Genomic distribution of RBM47 binding</i>	185
<i>Motif enrichment in RBM47 Binding sites</i>	188

<i>RBM47 re-expression in 831BrM2 cells leads to steady state levels changes of RBM47-bound transcripts</i>	189
<i>Independent validation of RBM47 mediated transcript level changes</i>	198
<i>Comparison of RBM47 bound and regulated targets to previously identified 831BrM2 signature expression changes</i>	198
<i>Validation of RBM47-mediated expression level changes in transcripts not previously identified as altered in metastasis</i>	206
<i>Analysis of paired-end RNAseq identifies extensive RBM47-dependent alternative splicing</i>	212
<i>RBM47 binds multiple transcripts in the region of alternative splicing associated with the epithelial to mesenchymal transition (EMT)</i>	219
<i>Independent validation of RBM47-regulated splicing changes</i>	219
<i>RBM47 alternative splicing RNA Binding Map</i>	226
Discussion	231
CHAPTER 6: General Discussion	
Summary	243
<i>Comparison of AGO and RBM47 HITS-CLIP binding maps reveal convergence on regulatory nodes of the Wnt pathway</i>	246
Future Directions	250
REFERENCES	253

LIST OF FIGURES

CHAPTER 3

Fig 3.1: <i>In vivo</i> selection of human breast cancer cell lines	49
Fig 3.2: Argonaute HITS-CLIP produces two data sets	54
Fig 3.3: Argonaute HITS-CLIP schematic	56
Fig 3.4: AGO HITS-CLIP in MDA-231 and derivative cell lines	57
Fig 3.5: Bioinformatic pipeline for processing AGO HITS-CLIP	59
Fig 3.6: Distribution of miRNAs bound to AGO	64
Fig 3.7: Definition of HITS-CLIP clusters and significant peaks	67
Fig 3.8: Genomic distribution of AGO-mRNA binding	70
Fig 3.9: Pairwise correlation of tags in AGO-mRNA clusters	72
Fig 3.10: Identification of <i>de novo</i> AGO binding sites – SPARC	74
Fig 3.11: AGO-mRNA binding sites gained in metastasis	78
Fig 3.12: Loss of AGO-mRNA binding correlates with HMGA1 steady state transcripts level	81
Fig 3.13: HMGA1 3'UTR AGO-mRNA differential binding	82
Fig 3.14: Seed sites in differentially bound HMGA1 cluster	84
Fig 3.15: HMGA1 3'UTR luciferase report assays reveal unexpected differential regulation	86
Fig 3.16: Differential AGO-mRNA binding in CD59 3'UTR	88
Fig 3.17: Differential AGO-mRNA binding correlates with CD59 mRNA and protein levels	90
Fig 3.18: CD59 is not regulated by miR-183	92

CHAPTER 4

Fig 4.1: Validation of pre-miR-335 overexpression in LM2 335OE	112
Fig 4.2: AGO-miRNA profiles in LM2 cell lines are unaffected by pri-miR-335 over expression	115
Fig 4.3: Overexpression of pre-miR-335 leads to an increase in AGO-miR335 and AGO-miR-335* in LM2 335OE cells	116
Fig 4.4: The stem-loop structure of human pre-miR-335	119
Fig 4.5: miR-335* is consistently detected at higher levels than miR-335	120
Fig 4.6: Genomic distribution of LM2 Control and LM2 335OE AGO HITS-CLIP clusters	122
Fig 4.7: AGO-mRNA binding is highly reproducible between LM2 Control and LM2 335OE cells	123
Fig 4.8: Experimental schematic of miR335/* dependent AGO-binding event identification	125
Fig 4.9: Non-canonical miR-335* seed sites in the AGO 62nt footprint in LM2 335OE enriched AGO clusters	130
Fig 4.10: RNAseq analysis of steady state transcript levels in LM2 Control and LM2 335OE cells	135
Fig 4.11: Differential AGO binding identifies CETN3 as a miR-335* target	138
Fig 4.12: CETN3 mRNA level is repressed by miR-335*	139
Fig 4.13: miR-335* represses the 3'UTR of CETN3	141
Fig 4.14: Endogenous miR-335* represses the CETN3 3'UTR	142
Fig 4.15: Investigating the role of CETN3 in invasion	144

Fig 4.16: Clinical association of CETN3 expression level and metastasis-free survival	145
Fig 4.17: Steady state mRNA levels of predicted miR-335/* targets show weak or no regulation by miR-335/*	147
Fig 4.18 Differential AGO binding identifies TWSG1 as a miR-335 target	149
Fig 4.19: SDAD1 and CDH11 are regulated at the protein level by miR-335* and miR-335	151
Fig 4.20: The DKK1 3'UTR shows evidence of two miR-335/* dependent binding events	154
Fig 4.21: Endogenous DKK1 transcript levels are strongly repressed in LM2 cells compared to MDA-231 and show no further repression by exogenous miR-335/*	155

CHAPTER 5

Fig 5.1: RBM47 expression is decreased in the highly metastatic derivatives the breast cancer cell line MDA-MB-231	170
Fig 5.2: RBM47 functions as a metastasis suppressor <i>in vivo</i> , with low expression clinically associated with poor metastasis-free survival	172
Fig 5.3: RBM47 gene and protein domain organization	174
Fig 5.4: Protein alignment of RBM47 with A1CF and hnRNP-Q	176
Fig 5.5: RBM47 is an RNA binding protein	178
Fig 5.6: Outline of the modified HITS-CLIP protocol used with RBM47	181

Fig 5.7: Summary of RBM47 HITS-CLIP sequencing data and reproducibility	184
Fig 5.8: Genomic distribution of RBM47 clusters	187
Fig 5.9: Motif enrichment in RBM47 binding sites	190
Fig 5.10: RNAseq analysis reveals RBM47-dependent steady state transcript level changes	192
Fig 5.11: RBM47 binding in 3'UTRs positively correlates with transcript up regulation	196
Fig 5.12: RBM47 binding does not correlate with transcript abundance	197
Fig 5.13: Dox-inducible Flag-RBM47 expression in clonal RBM47 WT#10 used for validation assays	199
Fig 5.14: SERPINB2 transcript levels are positively regulated by re- expression of RBM47 in 831BrM2	201
Fig 5.15: CTGF transcript levels positively correlate with RBM47 expression	202
Fig 5.16: RBM47-bound and downregulated genes show strong 3'end bias in the RNAseq profile	204
Fig 5.17: Cumulative distribution plots of genomic length of RBM47- bound genes showing RBM47 dependent transcript level changes	205
Fig 5.18: ANKRD1 transcript levels correlate with RBM47 expression	207
Fig 5.19: DKK1 transcript levels correlate with RBM47expression	208
Fig 5.20: RBM47 expression leads to increased stability of the DKK1 transcript	210

Fig 5.21: Normalized Ago-mRNA binding maps on RBM47-dependent upregulated transcripts	211
Fig 5.22: Overlap of RBM47 and Ago binding in 3'UTRs	213
Fig 5.23: Calculating change in inclusion rates of alternatively spliced isoforms from paired end RNAseq data	215
Fig 5.24: The majority of RBM47-dependent $\Delta I \geq 0.2$ alternatively spliced cassette exons show are bound by RBM47	216
Fig 5.25: Distribution of RBM47 unique tag numbers in BC2 \geq 2tags clusters	218
Fig 5.26: RBM47 binding correlates with RBM47-dependent exon exclusion in MDM4	222
Fig 5.27: RBM47 expression correlates with RBM47-dependent exon inclusion and exclusion in bound targets	224
Fig 5.28: Exogenous RBM47 expression correlates with RBM47-dependent exon inclusion and exclusion not seen when comparing MDA-2321 and 831BrM2	225
Fig 5.29: RBM47-RNA normalized complexity splicing map	227
Fig 5.30: No extended positional bias is evident in RBM47 binding flanking regulated alternative exons	230

CHAPTER 6

Fig 6.1: Multiple nodes of the DKK1 / Wnt pathway are regulated by RBM47 and AGO-miR-335/* with respect to metastatic potential	251
---	-----

LIST OF TABLES

CHAPTER 3

Table 3.1: Summary of AGO HITS-CLIP sequencing results from MDA- 231, 4175LM2, 1833BoM2 and 831BrM2	60
Table 3.2: Top 30 miRNA seed families in each cell line	62
Table 3.3: EdgeR analysis of AGO-miRNA HITS-CLIP reads	66
Table 3.4: AGO 3'UTR binding lost in all three metastatic sub-lines	80

CHAPTER 4

Table 4.1: Summary of AGO HITS-CLIP sequencing results from LM2 Control and LM2 335OE	113
Table 4.2: Analysis of the AGO-miRNA profiles of LM2 Control and LM2 335OE	118
Table 4.3: LM2 335OE 3'UTR clusters containing canonical seed matches in the 62nt AGO footprint	127
Table 4.4: LM2 335OE enriched significant peaks in CDS and 3'UTRs with canonical miR-335/* seed sequences	133
Table 4.5: miR-335/* canonical seed containing LM2 335OE-enriched AGO clusters downregulated in LM 335OE	136
Table 4.6: Previously published miR-335 target sites compared to LM2 335OE AGO HITS-CLIP	153

CHAPTER 5

Table 5.1: RBM47-bound transcripts showing RBM47-dependent increase in steady state transcript level	194
Table 5.2: RBM47-bound transcripts showing RBM47-dependent decrease in steady state transcript level	195
Table 5.3: RBM47-bound and dependent splicing changes previously identified as altered in EMT	220

CHAPTER 1: INTRODUCTION

Breast cancer is the most commonly diagnosed tumor in women, with an estimated ~230,500 newly diagnosed invasive cases expected in 2011, and is the second highest cause of female cancer-related mortality (American Cancer Society, 2012). Metastasis, a complex multistep process in which cells shed from the primary tumor disseminate throughout the body to colonize and proliferate in organs distant from the site of the original tumor, is the leading cause of death from breast cancer. The 5-year relative survival rate of patients presenting with distal metastases drops to 23.3% from 98.6% in patients with localized disease (Siegel et al., 2012). Dissemination of tumor cells can occur early in primary breast cancer progression leading to both primary tumor re-seeding and accelerated primary tumor growth (Kim et al., 2009b), and colonization of secondary metastatic niches leading to eventual outgrowth as distal metastatic lesions (Husemann et al., 2008). The metastatic process is inherently inefficient (Luzzi et al., 1998, Cameron et al., 2000) and is thought to require both tumor cell intrinsic adaptation and cell extrinsic metastatic niche modulation (Nguyen et al., 2009).

Metastatic cells show differing abilities to grow in certain organs with varying latencies to outgrowth that are dependent upon tumor type or sub-type, and in a way that cannot be solely be explained by circulatory patterns (Mendoza and Khanna, 2009, Nguyen et al., 2009). In breast adenocarcinoma, metastases are most commonly found in the auxiliary

lymph nodes, bone, lung, brain and liver (Hess et al., 2006, Patanaphan et al., 1988). Both general poor-prognosis and tissue-specific gene expression signatures have been described that clinically correlate with metastatic relapse (van 't Veer et al., 2002, van de Vijver et al., 2002, Bos et al., 2009, Kang et al., 2003, Li et al., 2011a, Minn et al., 2005a). However, understanding the exact molecular and cellular processes involved in both gain of metastatic ability and organ-specificity remains challenging (Sethi and Kang, 2011).

RNA Binding proteins and cancer-associated alterations in RNA metabolism

RNA metabolism is an integral part of both normal cellular complexity and the altered transcriptomes observed during both tumorigenesis and gain of metastatic ability. Central to this complexity are the concerted actions of multiple RNA binding proteins (RBPs) that regulate the spatial, temporal and functional dynamics of target RNAs from transcription to maturation, from modulation of content through to eventual degradation (Moore and Proudfoot, 2009). This includes (but is not limited to) alternative splicing, alternative polyadenylation and the regulation of transcript stability predominantly mediated through interactions in 3' untranslated regions (3'UTRs). Thus, RNA is no longer considered solely as a template for protein synthesis; rather its precise regulation is fundamental in the generation of biological complexity in both normal and disease states (Sharp, 2009, Licatalosi and Darnell, 2010).

Alternative Splicing

In eukaryotes precursor messenger RNA (pre-mRNA) is subjected to multiple post-transcriptional modifications including the intramolecular joining (splicing) of regions of coding sequence that will be transcribed into protein (exons), and the removal of intervening non-coding sequences (introns). Splicing reactions are catalyzed by a large multi-component ribonuclear complex called the spliceosome, which is comprised of five subcomponent small nuclear ribonucleoproteins (snRNPs U1-6) that assemble in a stepwise manner on core sequence elements within the intron to be removed. The sequences recognized by the spliceosome are relatively degenerate and in most cases the strength of the splicing signal is influenced by flanking pre-mRNA sequences, intronic or exonic splicing silencers or enhancers, which mediate their effects primarily by serving as binding sites for additional *trans*-acting regulatory RBPs (Wahl et al., 2009).

Alternative splicing was first described in 1980, when it was discovered that two protein products, one soluble and one membrane bound, could be produced from the same immunoglobulin gene (Alt et al., 1980, Early et al., 1980). Multiple types of alternative splicing have since been described; alternative 5' and 3' splice site selection, single exon (cassette-exon) inclusion or exclusion, mutually exclusive cassette-exon inclusion or exclusion, intron retention, and alternative promoter or 3'exon selection (Keren et al., 2010). Recent studies using high-throughput sequencing of the human transcriptome indicate that alternative splicing

is major contributor to proteomic diversity, with over 90% of human genes showing evidence of alternative transcripts (Pan et al., 2008, Wang et al., 2008). Additionally, patterns of tissue-specific regulation have been observed indicating that the enormous diversity of proteins that can potentially be generated by alternative splicing is regulated in a coordinated manner (Wang et al., 2008). For example, the neuronal specific splicing factor Nova has been found to regulate the alternative splicing of a biologically coherent set of transcripts involved in synaptic plasticity, providing the first demonstration of an RBP-coordinated biological network (Ule et al., 2003, Ule et al., 2005b, Licatalosi et al., 2008).

The regulation of alternative splicing is dependent upon interactions between multiple RBPs, and between splicing factor RBPs and their binding motifs in target pre-mRNAs (Licatalosi and Darnell, 2010). It has been shown that splicing factors can be differentially expressed between human tumors and normal tissues, although for most splicing factors it is not known whether these tumor specific expression patterns are a consequence or causative of the oncogenic process (Kirschbaum-Slager et al., 2004). It is interesting to note that alternative splicing patterns can be used to distinguish normal tissue from serous ovarian tumors (Venables et al., 2009), breast tumors (Venables et al., 2008) and colon cancers (Gardina et al., 2006) suggesting that alteration in alternative splicing is a common feature of tumorigenesis.

Cancer-associated mutations resulting in alternative splicing

Germline mutations in high penetrance susceptibility genes such as BRCA1, BRCA2, PTEN and TP53 contribute approximately 20% of familial genetic risk in breast cancer (Thompson and Easton, 2004), with BRCA1 and BRCA2 mutations conferring a lifetime risk of 80-90% (Meindl et al., 2011). The BRCA proteins are tumor suppressors involved in homologous recombination and error-free repair of DNA double strand breaks (Roy et al., 2012). Interestingly, previously unclassified germline variants of BRCA1/2 associated with breast cancer, those that do not have predicted effects on protein function, are being characterized as anomalous splicing variants resulting from disruption of regulatory splicing elements (Sanz et al., 2010, Wappenschmidt et al., 2012) or by activation of cryptic exons (Anczukow et al., 2012a).

It is estimated that 10% of human inherited diseases result from mutations disrupting components of 5'- or 3'-splice sites (Cooper et al., 2009). However, this estimate does not take into consideration exonic mutations as these are generally presumed to result in missense mutations and altered protein function, or to be neutral due synonymous mutations. Exonic mutations have the potential to create or destroy regulatory sequences known as exonic splicing enhancers (ESE) or silencers (ESS). Recent bioinformatic analyses of exonic disease-causing human mutations have found a strong association between single nucleotide polymorphisms (SNPs) and both the creation of ESS motifs and ablation of ESEs suggesting that exon skipping may play a prominent role in human

inherited disease (Sterne-Weiler et al., 2011, Lim et al., 2011). Furthermore, trait associated SNPs, mutations significantly and reproducibly associated with but not causative of particular inherited diseases, are also enriched for exon-exclusion inducing intronic mutations (Lee et al., 2012) suggesting that mis-splicing is a common feature of genetic disease.

Large-scale changes in splicing patterns have been observed in multiple tumor types during the transition from normal to cancerous tissue, making widespread germline or gain of tumor-associated mutations highly unlikely. Mutations in proteins involved in the core splicing machinery have been found in myelodysplasia, a neoplasm that predisposes for acute myeloid leukemia (Yoshida et al., 2011), however deregulation of trans-acting alternative splicing modulators are more commonly described.

Splicing regulators as oncogenes – SRSF1

The SR protein family members, serine/arginine rich splicing factor (SRSF) 1-12, are defined based on the presence of 1-2 RNA recognition motif (RRM) domains followed by an ~50nt >40% serine/arginine domain of consecutive RS or SR repeats (Manley and Krainer, 2010). SR proteins are involved in both constitutive and alternative splicing of pre-mRNA through binding to ESEs (Busch and Hertel, 2012). Karni et al. (2007) discovered that SRSF1 (previously known as SF2/ASF) is upregulated in lung, colon and breast tumors when compared to normal tissues, and that ectopic expression of SRSF1 is

oncogenic both *in vitro* and *in vivo*. Over-expression of SRSF1 leads to the alteration of alternative splicing patterns of endogenous target genes, including increased expression of an oncogenic isoform of RPS6KB1, and increased inclusion of an alternative exon12a of BIN1 that abolishes its tumor suppressor activity (Karni et al., 2007).

Subsequently it has been shown that SRSF1 is transformative in mammary epithelial cells (Anczukow et al., 2012b), and is upregulated in lung adenocarcinoma and squamous cell lung cancer (Gout et al., 2012), non-small cell lung cancer (Ezponda et al., 2010), and pediatric acute lymphoblastic leukemia (Zou et al., 2012). SRSF1 regulates the inclusion of exon 3b of Rac1 in colon cancer cells (Goncalves et al., 2009) resulting in the expression of a constitutively active Rac1b isoform that has been shown to promote transformation of NIH3T3 cells through AKT phosphorylation and activation (Singh et al., 2004). Additionally, the transforming ability of SRSF1 in immortal fibroblasts has been demonstrated to be dependent upon mTORC activity (Karni et al., 2008), while SRSF1 expression is both directly regulated by, and necessary for, the oncogenic activity of Myc in lung cancer cells (Das et al., 2012). A genome-wide binding map of SRSF1 has been generated in mouse embryonic stem cells using high-throughput cross-linked immunoprecipitation (HITS-CLIP) which, in conjunction with splicing-sensitive microarray analysis, has identified 498 SRSF1-dependent alternative splicing events (Pandit et al., 2013), although it is not known

how many of these targets are expressed and regulated in the tumor tissues associated with SRSF1 transformation.

Splicing regulators as oncogenes – FOX2

Using genome-wide, high throughput RT-PCR, Venables et al. (2009) identified multiple alternative splicing events common to tumors of the breast and ovary that can discriminate between normal and cancer tissues. Analysis of sequences in the 100nt flanking regions of these cancer associated-exons revealed enrichment of FOX1/2 binding motifs. FOX1 mRNA is detected almost exclusively in the brain and striated muscle, while FOX2 is more broadly expressed in heart, brain and the ovary (Jin et al., 2003, Lieberman et al., 2001, Underwood et al., 2005). Subsequent analysis of breast cancer and non-malignant immortalized breast epithelial cell lines have revealed that FOX2 regulates alternative splicing patterns associated with claudin-low, basal breast cancer sub-types (Lapuk et al., 2010). Computational analyses of the genome-wide distribution of FOX1/2 binding motifs have revealed a position-dependent relationship between alternative exon inclusion or exclusion (validated by semi-quantitative RT-PCR) in transcripts enriched in gene ontology terms associated with cytoskeletal organization and neuromuscular functions (Zhang et al., 2008). Genome-wide mapping of FOX2 on target mRNAs in human embryonic stem cells (hECS) has confirmed this positional bias in splicing and identified FOX2 as necessary for hECS survival, although this phenotype was not found in 293T cells or

human fetal neural stem cells suggesting cell type specific FOX2 target regulation (Yeo et al., 2009).

Splicing regulators as tumor suppressors – RBM5

Allele loss at 3p21.3 is both the earliest pre-malignant change thus far detected in lung cancer (Sundaresan et al., 1992) (Hung et al., 1995, Sundaresan et al., 1992, Wistuba et al., 1999), and the most frequent lung cancer-associated genetic alteration (Zabarovsky et al., 2002). This ~630kb deletion is also observed in multiple other tumor types including breast, and spans multiple putative tumor suppressor genes including the RNA binding protein RBM5 (Lerman and Minna, 2000, Zabarovsky et al., 2002). RBM5 encodes two RRM domains, a zinc finger motif, a bipartite nuclear signal and a G-patch domain (Timmer et al., 1999). Unlike the pro-tumorigenic splicing factors mentioned above, RBM5 is tumor suppressive; inhibiting growth in MCF-7 breast cancer cells (Oh et al., 2002), A549 non-small cell lung cancer (NSCLC) cells (Oh et al., 2006, Shao et al., 2012) and CEM-C7 leukemia cells (Mourtada-Maarabouni et al., 2003).

RBM5 is reported to promote apoptosis (Mourtada-Maarabouni et al., 2003, Shao et al., 2012) in part through binding to and positively regulating the splicing of a pro-apoptotic isoform of caspase 2, casp-2L (Fushimi et al., 2008). Conversely, RBM5 can interfere with a late step in spliceosome assembly, promoting exon exclusion and production of a non-membrane-bound Fas molecule that inhibits apoptosis (Bonnal et al.,

2008, Cheng et al., 1994). Additionally, RBM5 causes cell cycle arrest in G1 in lung cancer cells, and induces the expression of the apoptotic activator BCL2-associated protein X (Bax) although the mechanisms by which this occurs remain unknown (Oh et al., 2006).

Alternative splicing in metastasis

Multiple transcripts that promote tumor cell invasive behavior are regulated by alternative splicing (David and Manley, 2010). The transmembrane protein CD44 is encoded by 20 exons, 10 of which can be regulated by alternative splicing leading to multiple different functional proteins (Ponta et al., 2003). Alternative splice isoforms of CD44 were among the first to be associated with induction of metastasis (Gunthert et al., 1991) and have subsequently been found to promote metastasis in multiple tumor types. Regulation of CD44 splicing is reported to be controlled by multiple candidate RBPs including SFRS10 (Watermann et al., 2006, Takeo et al., 2009), PCBP-1/hnRNP-E1 (Zhang et al., 2010b, Lian et al., 2012), Sam68 (Matter et al., 2002) in association with SRm160 (Cheng and Sharp, 2006), and ESRP1 (Reinke et al., 2012, Yae et al., 2012). While the majority of these RBPs enhance inclusion of various combinations of alternative CD44 exons, PCBP-1/hnRNP-E1 promotes exon exclusion suggesting that over-all splicing outcome in any given cellular context depends upon the regulatory proteins expressed.

Altered patterns of alternative splicing have been observed when comparing primary tumors with differing metastatic capacity (Dutertre et

al., 2010), and during the acquisition of the invasive phenotype that accompanies the epithelial-mesenchymal transition (EMT) (Warzecha et al., 2010, Shapiro et al., 2011). The EMT involves loss of cell polarity and cell-cell junctions and the acquisition of a more motile, invasive cellular phenotype via extensive transcriptional changes and reorganization of the actin cytoskeleton. While direct evidence linking EMT to metastasis *in vivo* is difficult to obtain, the same processes commonly involved in developmental EMT bear striking similarity to those involved in metastasis (Yilmaz and Christofori, 2009).

Three RBPs are currently implicated as major regulators of EMT-associated alterations in alternative splicing patterns: ESRP1 and ESRP2, and RBFOX2 (Shapiro et al., 2011). The switch between the mutually exclusive inclusion of fibroblast growth factor receptor 2 (FGFR2) exon3b (epithelial) or exon3c (mesenchymal) that occurs during EMT and metastasis (Yan et al., 1993) is thought to be a direct result of downregulation of the epithelial specific splicing factors ESRP1 and ESRP2 (Warzecha et al., 2009). Knockdown of ESRP1 and 2 in mammary epithelial cells induces both the cellular changes indicative of EMT and extensive alterations in splicing of transcripts associated with vesicular transport (Warzecha et al., 2010), while re-expression of ESRP1 in mesenchymal cells causes partial reversion to the epithelial phenotype (Shapiro et al., 2011). Analysis of RBP expression levels in mammary epithelial cells undergoing EMT have confirmed both the loss of ESRP1/2, and identified a slight up regulation of RBFOX2 (Shapiro et al., 2011) a

known driver of mesenchymal-specific splicing patterns (Venables et al., 2013). Depletion of RBFOX2 in mesenchymal cells leads to partial reversion to the epithelial phenotype, while EMT associated up-regulation of RBFOX2 leads to altered splicing patterns of predicted targets associated with invasion (Braeutigam et al., 2013).

RNA binding proteins and modulation of transcript stability

The relative abundance of mRNA is a function both of the rate of transcription and degradation, with each altered in response to stimuli in order to adjust the resulting protein level. Aberrant gain or loss of a transcript in disease may reflect either DNA mutation or dysregulation of trans-activating transcription factors, or could result from altered post-translational regulation of stability of the message. mRNA stability can be regulated by RBPs, either alone or in complex with non-coding RNAs, through binding to *cis*-regulatory elements within the transcript (Wu and Brewer, 2012). Several well characterized mRNA stabilizing and destabilizing factors are known to be deregulated in tumorigenesis and metastasis, influencing the establishment of aberrant target mRNA expression profiles.

Nonsense mediated decay in cancer

Nonsense mediated decay (NMD) is a translation-coupled mRNA surveillance process that selectively eliminates transcripts containing premature stop codons to minimize synthesis of truncated proteins. During splicing of pre-mRNA, the exon junction complex (EJC) is

deposited approximately 20-24 nucleotides upstream of the exon-exon junction, providing a position dependent memory of the splicing event (Le Hir et al., 2000, Le Hir et al., 2001). If a translating ribosome encounters a stop codon >50-55nt upstream of an EJC, translational pausing occurs leading to the recruitment of the SMURF complex (SMG1, UPF1 and eukaryotic release factors) and SMG1 phosphorylation of UPF1 (Kashima et al., 2006). This in turn leads to the recruitment of decapping enzymes, the endoribonuclease SMG6 and deadenylation, resulting in the rapid decay of the mRNA (Lejeune et al., 2003, Huntzinger et al., 2008).

Meta-analysis of predicted human disease-causing mutations has revealed a disproportionately high number of nonsense mutations in tumor suppressor genes when compared to missense mutations, with the opposite being true for oncogenes (Mort et al., 2008). NMD has been shown to prevent the translation of truncated proteins in breast cancer that arise from germline mutations in BRCA1, CHK2 and TP53 (Perrin-Vidoz et al., 2002, Anczukow et al., 2008). In hereditary diffuse gastric cancer, ~80% of the causative germline mutations in e-cadherin generate premature stop codons leading to NMD and transcript down regulation (Karam et al., 2008). In addition to DNA mutation, introduction of premature stop codons can also be achieved through alternative splicing, by inclusion of occult, poorly conserved exons that result in frame-shift mutations (Green et al., 2003, Lewis et al., 2003).

While it has been argued that the inclusion of the majority of mammalian alternatively spliced NMD exons results from “noisy” splicing rather than regulated events (Zhang et al., 2009), the onconeural antigen proteins Nova1 and Nova2 have been shown to bind and regulate the alternative splicing of multiple NMD exons in brain in an activity dependent manner (Eom et al., 2013). This would suggest that inclusion of NMD exons and hence down regulation of target mRNA can proceed in a coordinately regulated manner. The NOVA proteins were first identified as the target antigens in paraneoplastic opsoclonus-myoclonus ataxia (POMA) syndrome (Darnell et al., 1991), a neurological degenerative disorder commonly associated with ectopic NOVA expression in breast and gynecological cancers (Luque et al., 1991). Ectopic NOVA expression is also reported in Hodgkin lymphoma cells (Religio et al., 2005) and in primary serous ovarian tumors (Darnell lab, unpublished data). Additionally, multiple splicing factors have been shown to be auto- or trans-regulated via inclusion or exclusion of NMD exons. For example, PTBP1-dependent skipping of exon11 in PTBP2 leads to a frame-shift mutation that targets PTBP2 mRNA for NMD and establishes mutually exclusive expression patterns of the two splicing factors (Rahman et al., 2002, Boutz et al., 2007). It has been discovered that NMD is inhibited in the tumor microenvironment leading to enhanced tumorigenesis (Wang et al., 2011b) that is dependent upon c-myc activity (Wang et al., 2011a). Since many NMD isoforms are expressed at extremely low levels, the extent of NMD-associated transcript regulation by RBPs in tumorigenesis and metastasis may currently be underestimated.

Transcript stability, ARE-binding proteins and cancer

AU-rich elements (AREs) are 50-150nt sequences rich in adenosine and uracil bases predominantly found in the 3'UTRs of mRNAs with short half-lives, including those involved in transient biological processes such as immune responses, cellular growth and external stress-mediated pathways (Bakheet et al., 2001). AREs were first identified as destabilizing sequences when it was found that fusing the AT-rich 3'UTR segment of the unstable granulocyte-macrophage colony stimulating factor (GM-CSF) to the normally stable rabbit globulin transcript induced rapid degradation (Shaw and Kamen, 1986). Many of the over twenty ARE-binding proteins identified to date have been shown to promote target instability by recruiting mRNA decay proteins to their targets (Wu and Brewer, 2012). Among this destabilizing cohort is AUF1/hnRNP-D, a family of four proteins generated from a single gene through alternative splicing (Wagner et al., 1998) that interact with AREs via two tandemly arranged, non-identical RRM domains (DeMaria et al., 1997). AUF1 has been shown to interact with and destabilize the mRNA of multiple cancer related proteins, including Bax and junD (Mazan-Mamczarz et al., 2008b, Zou et al., 2010, Zucconi and Wilson, 2011).

Conversely, direct interactions with ARE-binding proteins can lead to stabilization of targets, for example, binding by the ELAVL family of proteins (ELAVL1/HuR, nELAV: HuB/C/D) generally stabilizes mRNAs bearing AREs and U-rich sequences. The neuronal ELAVL family members (nELAV) (Okano and Darnell, 1997) were discovered as the

auto-antigens in paraneoplastic encephalomyelitis (PEM) (Szabo et al., 1991), a neurological disorder associated with small cell lung cancer (SCLC) (Dalmau et al., 1990, Manley et al., 1995). PEM is one of a diverse group of paraneoplastic neurologic disorders (PNDs) characterized by high-titer antibodies recognizing neuronal antigens that are expressed in occult tumors. Patients with PEM have high titers of anti-nELAVL antibodies (anti-Hu, which were used to clone the cDNAs encoding HuC and HuD (Szabo et al., 1991)), but while all SCLCs appear to express nELAVL antigens (Budde-Steffen et al., 1988, Dalmau et al., 1995) the neurological syndrome is rare. HuD expression has also been associated with neuroblastoma (Dalmau et al., 1992, Ball and King, 1997) where it is believed to stabilize N-myc (Lazarova et al., 1999), with loss of heterozygosity of HuD associated with N-myc gene amplification (Grandinetti et al., 2006). Upregulation of the ubiquitously expressed ELAVL1 is associated with multiple cancers including colon (Lopez de Silanes et al., 2003), lung (Blaxall et al., 2000), and breast (Denkert et al., 2004, Heinonen et al., 2005, Mazan-Mamczarz et al., 2008a, Mazan-Mamczarz et al., 2008b). The repertoire of ELAVL-family regulated transcripts has been greatly expanded by use of RNA cross-linking methodologies (Keene et al., 2006, Mazan-Mamczarz et al., 2008a, Mukherjee et al., 2011, Lebedeva et al., 2011, Ince-Dunn et al., 2012 and Darnell lab unpublished data), with many of the identified targets implicated in the acquisition of cancer-related phenotypes (Simone and Keene, 2013).

MicroRNA mediated control of gene expression in oncogenesis and metastasis

MicroRNAs (miRNA) are a major class of evolutionarily conserved, small (~22nucleotide (nt)) non-coding RNA (ncRNA) molecules that are associated with diverse biological processes including proliferation, differentiation, and stress response through post-transcriptional regulation of target mRNAs. miRNA are transcribed by either RNA polymerase II or III from intergenic regions of the genome, or as introns in the pre-mRNA of coding and ncRNAs, and can be transcribed either as the sole miRNA in that region or as part of a co-regulated cluster. miRNA begin as longer RNA molecules (primary or pri-miRNA) that are cleaved in the nucleus by the RNase III endonuclease Drosha to liberate a ~60-70nt stem loop with one defined mature miRNA end. This pre-miRNA is actively exported out of the nucleus by Ran-GTP and Exportin-5, and processed by a second RNase III endonuclease, Dicer. Dicer recognizes the double stranded portion of the pre-miRNA hairpin structure and cuts at a fixed distance to generate a double stranded imperfect duplex with a characteristic ~2nt 3'-overhang. One of these strands (mature miRNA) is loaded into Argonaute (AGO), the core component of the miRNA-induced silencing complex (miRISC). The second, passenger strand (miRNA*) is generally found at lower levels than the mature miRNA and is presumed to be degraded (Bartel, 2004). Complementarity between the 5' end of the mature miRNA in positions 2-8, or seed region, targets the AGO-miRNA complex predominantly to 3'UTRs of target mRNA. AGO-ternary complex binding leads to translational repression and/or target mRNA destabilization (Filipowicz et al., 2008), although miRNA mediated

upregulation of targets has also been described under certain cellular conditions (Vasudevan, 2012).

A general downregulation of miRNAs has been observed in tumors, with patterns of individual miRNA expression being highly informative for both classification and prognosis (Lu et al., 2005, Calin et al., 2005, Rosenfeld et al., 2008). Such general downregulation may be achieved by transcriptional repression of multiple miRNAs by oncogenic transcription factors such as Myc (Chang et al., 2008), by loss of p53 transactivation (Suzuki et al., 2009), or downregulation/ altered activity of the miRNA biogenesis pathway (Thomson et al., 2006, Kumar et al., 2007). Haploinsufficiency of Dicer has been detected in multiple human tumors (Hill et al., 2009, Kumar et al., 2009), with loss of a single allele leading to accelerated tumor formation *in vivo* (Kumar et al., 2009, Lambertz et al., 2010).

In addition to being regulated by oncogenes and tumor suppressors, individual miRNA or miRNA clusters can themselves be either pro- or anti-tumorigenic depending on cellular context. Greater than 50% of miRNAs are located in fragile chromosomal regions that frequently undergo deletion and amplification in human tumors (Calin et al., 2004, Zhang et al., 2006), and many can be epigenetically regulated (Choudhry and Catto, 2011, Scott et al., 2006). Loss or gain of a single miRNA can have a profound effect on cellular physiology due to the concordant regulation of potentially hundreds of target mRNAs

(Hendrickson et al., 2009, Lim et al., 2005). Multiple miRNAs have been implicated in the development of breast cancer as both tumor suppressive and oncogenic factors (Spizzo et al., 2009, Tang et al., 2012a). For example, upregulation of miR-155 has been found in multiple breast cancer subtypes where it is associated with more aggressive tumors and poor prognosis (Mattiske et al., 2012). Overexpression of miR-155 in breast cancer cells leads to enhanced proliferation and survival, while deletion sensitizes cells to apoptosis and chemotherapeutic reagents (Jiang et al., 2010, Kong et al., 2010). Interestingly, wild-type BRCA1 function has been associated with the epigenetic silencing of miR-155 through promoter region binding and recruitment of HDAC2 (Chang et al., 2011). Conversely, loss of the tumor suppressive miR-205 promotes anchorage-independent breast cancer cell growth (Wu et al., 2009b), in part through derepression of HER3 tyrosine-protein kinase receptor and subsequent increased activation of the PI3K/ AKT survival pathway (Iorio et al., 2009).

Recent studies have identified several miRNAs with roles in both metastasis suppression and promotion. miR-10b is upregulated in metastatic breast cancer cells compared to immortalized human mammary epithelial cells (hMECs), with knockdown reducing the invasive ability of MDA-MB-231 cells and ectopic expression enhancing the invasion and migration of weakly metastatic cells both *in vitro* and *in vivo* (Ma et al., 2007). miR-10b expression is positively associated with high-grade malignancy of various cancer types, with expression shown to be elevated in the primary breast tumors of metastasis-positive patients

(Baffa et al., 2009). Systemic treatment with a miR-10b antagonist has been shown to suppress lung metastasis in a mouse breast cancer model if given during early stages of tumor cell dissemination (Ma et al., 2010b). In contrast, down regulation of multiple endogenous miRNAs has been shown to promote metastasis in breast cancer. Loss of miR-335, miR-126 and miR-206 promotes a highly aggressive metastatic breast cancer phenotype both *in vivo* and *in vitro* (Tavazoie et al., 2008), while re-expression of miR-31 both inhibits breast cancer metastasis (Valastyan et al., 2009) and can elicit regression of already established lung metastases (Valastyan et al., 2011). While some progress has been made in identifying targets of these pro- and anti-metastatic miRNAs, current studies have been limited by lack of knowledge of direct sites of AGO-mRNA binding and by restricting target searches to only those transcripts that both contain bioinformatically predicted or conserved seed sites and show large miRNA-dependent downregulation.

Alternative polyadenylation site selection

With the exception of some histone transcripts, all eukaryotic mRNAs undergo a two-step maturation process involving endonucleolytic cleavage and subsequent untemplated polyadenine (polyA) tail addition. The specificity and efficiency of 3' end processing is dependent upon a multiprotein complex binding to *cis*-acting sequence elements in the pre-mRNA known as the polyA signal (Danckwardt et al., 2008). Approximately 50% of human transcripts exhibit conserved alternative polyadenylation site usage as a result of tandem polyA sites,

polyA sites within the terminal intron, alternative splicing of the terminal exon, or polyA site usage within an internal exon (Tian et al., 2005, Yan and Marr, 2005). While the intrinsic strength of the polyA site sequence and the concentration/activity of polyadenylation factors can together determine alternative polyA site usage, various other tissue or stage-specific RBPs can also influence polyA site choice (Barabino and Keller, 1999). For example, alternative polyA site usage can be regulated by Nova binding in 3'UTR regions flanking polyA sites (Licatalosi et al., 2008).

Use of alternative polyA sites has the potential to alter transcript stability via elimination of large stretches of the 3'UTR, enabling escape from any negative *cis*-regulatory elements located in the longer 3'UTR isoform. Systematic analysis of human and mouse transcriptomes have shown that shorter 3'UTR isoforms are most abundant when transcripts are expressed at high levels, while longer 3'UTRs are associated with lower expression levels (Ji et al., 2011). Oncogenic chromosomal rearrangements in benign tumors that involve the loss of the c-terminal coding sequence and entire 3'UTR of HMGA2 (Geurts et al., 1997, Schoenmakers et al., 1995) have been shown to result in relief of let-7-mediated repression (Mayr et al., 2007). Widespread shortening of 3'UTRs with subsequent loss of miRNA regulatory sites has been described both in activated T cells compared to resting T cells, and as a general feature of proliferating cells (Sandberg et al., 2008). Furthermore, cancer cell lines have been found to express shorter 3'UTR isoforms than non-transformed cells, with 3'UTR shortening leading to activation of oncogenes (Mayr and

Bartel, 2009). The mechanism (or RBPs responsible) for the phenomena of alternative polyA site selection and global shortening of 3'UTR length in oncogenic activation remains unexplored.

Mapping the transcriptome-wide regulatory networks of RNA binding proteins

Evidence of functional roles for RBPs in tumorigenesis and regulation of metastasis is extensive. In the studies outlined above, a reductionist approach is commonly taken to identify one or two targets that can partially explain the pro- or anti-metastatic consequences of altered RBP expression. However many RBPs are multifunctional and can regulate multiple steps in RNA metabolism. For example, the onconeural Nova proteins can regulate transcript content via alternative splicing, and stability through regulation of NMD and alternative polyA site usage. The ELAVL proteins while generally recognized as ARE-binding modulators of transcript stability also have roles in alternative splicing (Ince-Dunn et al., 2012, Lebedeva et al., 2011). Additionally, the FET family of RBPs (FUS/TLS, EWSR1 and TAF15) associated with chromosomal rearrangements in Ewing's sarcoma, a highly aggressive primary tumor of the soft tissue and bone with a high propensity to metastasize (Sankar and Lessnick, 2011), have multiple roles in RNA metabolism, including regulation of transcription (Bertolotti et al., 1996, Bertolotti et al., 1998, Lagier-Tourenne et al., 2012), alternative splicing (Paronetto et al., 2011, Ishigaki et al., 2012, Lagier-Tourenne et al., 2012, Rogelj et al., 2012,

Ibrahim et al., 2013) and the biogenesis of miRNAs (Gregory et al., 2004, Ballarino et al., 2012, Morlando et al., 2012).

Given the potential for complex patterns of RNA regulation, comprehensive evaluation of directly regulated transcripts is necessary to fully understand the downstream mediators of the anti- or pro-metastatic activities of any given RBP or miRNA. While it is possible to predict targets via observations of genome-wide changes in transcript levels upon RBP/miRNA modulation in combination with bioinformatic assessment of associated *cis*-elements (where consensus regulatory sequences are known), this type of analysis does not discern direct versus indirect effects. This is particularly pertinent when considering regulation by RBPs that have the potential to influence the activity of multiple other RBPs, for example, PTBP1 down regulation of PTBP2 via alternative splicing. These limitations also apply to AGO-mediated miRNA regulation of transcripts as bioinformatic predictions of regulatory sites suffer from high false positive rates even when taking conservation and extended regions of seed match into consideration (Baek et al., 2008, Selbach et al., 2008). Furthermore, reporter assays may produce false-positive or false-negative results if not carefully designed to include the native 3'UTR context, particularly when considering the potential for regulated alternative polyA site usage.

Multiple techniques have been developed to biochemically identify functional RBP-RNA interactions *in vivo*. Immunoprecipitation of an RBP

under mild conditions coupled with purification of co-precipitating RNA, either in the presence or absence of formaldehyde crosslinking, allows for coarse mapping of an RBP on target RNA (Keene et al., 2006). However this technique suffers from low signal-to-noise, has poor binding site resolution, and favors capture of the most stable RBP-RNA interactions (Mukherjee et al., 2011). Additionally, this technique does not discriminate between direct RBP-RNA interactions and those occurring due to either co-immunoprecipitation of other RBPs or random RNA-RBP reassociation post-lysis (Mili and Steitz, 2004, Riley et al., 2012).

Direct mapping of RNA-protein complexes became possible with the development of crosslinking immunoprecipitation (CLIP) strategies (Ule et al., 2003, Ule et al., 2005a, Jensen and Darnell, 2008). UV-irradiation induces covalent bond formation between closely opposed RNA and protein molecules within physiologically relevant distances (on the order of Angstroms apart) but does not induce protein-protein crosslinks (Darnell, 2010). Such UV-induced covalent bonds allow for rigorous purification of RBP-RNA complexes from live cells or whole tissues, which after reverse transcription and sequencing yield a population of cDNA clones corresponding to the regions of RNA directly bound by the protein of interest. By coupling CLIP to next generation sequencing (high-throughput CLIP, HITS-CLIP (Licatalosi et al., 2008)), high-resolution unbiased transcriptome-wide footprints of RNA-protein interactions can be generated. Furthermore, with the growing number of HITS-CLIP binding maps now available from different tissues (brain, stem cells,

tissue culture) and multiple RBPs with known RNA metabolic activities, it is possible to predict functional outcomes from binding patterns on target transcripts with no prior knowledge of an RBP of interests function. For example, robust binding in intronic regions has been associated with the regulation of alternative splicing by multiple RBPs, including but limited to Nova (Licatalosi et al., 2008, Ule et al., 2003, Ule et al., 2006, Ule et al., 2005b), the ELAVL proteins (Ince-Dunn et al., 2012, Lebedeva et al., 2011, Mukherjee et al., 2011), and FUS (Ishigaki et al., 2012, Lagier-Tourenne et al., 2012, Nakaya et al., 2013, Rogelj et al., 2012). Binding sites in 3'UTR regions have been used to both delimit the functional regulatory sites of AGO and to identify the targeting miRNA through narrowed search for complementary seed sites within these footprints (Chi et al., 2012, Loeb et al., 2012, Zisoulis et al., 2010, Hafner et al., 2010).

Perspectives

HITS-CLIP provides a tractable methodology with which to study both RBP regulated events in tumor cells, and by differential analysis, to identify altered sites of regulation associated with gain in metastatic potential. To our knowledge HITS-CLIP has not been used to interrogate the targets of endogenous RBPs in breast cancer cells, nor has it been used to investigate RBP-associated regulation of metastasis. Although several cancer-related RBPs have been interrogated by HITS-CLIP, none of these experiments were carried out in cancerous cells or tissue, therefore the relevance of identified targets to cancer biology is unknown. Given the important dual role of miRNAs in the positive and negative regulation of

metastasis and the potential for the regulation of hundreds of targets (Hendrickson et al., 2009, Lim et al., 2005), AGO HITS-CLIP provides an attractive method by which the entire network of miRNA regulation in a cancer cell can be mapped. In this study we have generated HITS-CLIP binding maps of endogenous AGO in weakly and highly metastatic breast cancer cells leading to the identification of altered AGO association with both miRNA and target mRNAs with respect to metastatic ability and tissue tropism. Furthermore, we have assessed the genome-wide regulatory network of miR-335, a known anti-metastatic miRNA. Additionally, we have used HITS-CLIP as general platform for discovery of transcriptome-wide regulation by the previously uncharacterized, putative metastasis suppressor RBM47.

CHAPTER 2: EXPERIMENTAL PROCEDURES

Cell Culture

Triple reporter MDA-MB-231, 4175LM2, 1833BoM2 and 831BrM2 cell lines were obtained from the Massagué laboratory at MSKCC. Triple reporter 4175LM2 shID and 4175LM2 miR-335 over expressing cells, plus non-reporter MDA-MB-231 and 4175LM2 cells were obtained from the Tavazoie lab at Rockefeller University. All MDA-231 and derivative cell lines were cultured in D10F - DMEM (Corning) supplemented with 10% FBS (Hyclone), 1% sodium pyruvate, 1% penicillin-streptomycin and 0.4% fungizone (Gibco). HEK293T cells were obtained from ATCC and cultured in DMEM with 10% FBS. Non-clonal dox-inducible 831BrM2 triple reporter stable cell lines expressing WT FLAG-tagged RBM47 or Tet-On (empty vector) and the clonal 831BrM2 WT FLAG-tagged RBM47#10 cell line were generated by Sakari Vanharanta (Massagué Laboratory, MSKCC). RBM47 expression was induced using 1ng/ml or 10ng/ml doxycycline (Sigma) respectively for three days. All cells were maintained at 37°C in 5% CO₂.

Analysis of miRNA and mRNA expression levels by quantitative-PCR

Total RNA was purified by Trizol extraction (Invitrogen) followed by ethanol precipitation, DNase treatment (RQ1 RNase-free DNase (Promega) and phenol-chloroform extraction. 90ng of total RNA was used for reverse transcription with the miScript II RT Kit (Qiagen) followed Q-PCR in using the equivalent of 2ng starting RNA with FastStart SYBR

Green Master mix (Roche Applied Science) and commercially available miScript miRNA specific primers (Qiagen) as per the miScriptII RT protocol. Primers specific to SNORD44 were used as a normalization control. For mRNA expression analysis, 1µg of total RNA was reverse transcribed using the iScript cDNA synthesis kit and 100ng used per FastStart SYBR reaction. Genom analysis (Vandesompele et al., 2002) was carried out using the BioGazelle software on MDA-231 and its metastatic derivatives to select a control gene for normalization. TATA-binding protein (TBP) showed the least variation between cell lines and was used as the control gene for all mRNA Q-PCR. For all analyses, Q-PCR was done in triplicate for each independently prepared RNA sample (minimum of three per condition/cell line) using a BioRad IQ5 Real Time PCR machine and 58°C annealing temperature, with both melt curves and -RT controls performed. All Q-PCR primers were designed using Primer3Plus.

HMGA1 F – agcgaagtgccaacacctaa	R – ttccttctctggagttgtgg
CD59 F - cacaatgggaatccaaggag	R - tgcagtcagcagttgggtta
CETN3 F – tggacaaaacaaagaggaaaaa	R – acatcaaaccceaaggctct
SEL1L F – agaaggcagccaggatgaat	R – gaatcaagaaatatttgaccagca
FAM105B F – agcaatggaaacttgactg	R – tttgtgaatagttcatcacaagcta
SMAD5 F – tctgcttgggtttgtgtca	R – ctgctgtcactgaggcattc
TWSG1 F – aatgttcacgcgccttattc	R - aaccagcgatatttgatgc
DKK1 F – gatcatagcaccttggatggg	R – ggcacagtctgatgaccgg
RBM47 F – ctgtcattcccactgtgtcg	R - gtaactggccccgtagatcc

SERPINB2 F – ggtcctggtgaatgctgtct	R - ctgtacaggtgtgctgag
CTGF F – ggagtgggtgtgtgacgag	R – ccaggcagttggctctaac
ANKRD1 F – gagataccccgttgcacgat	R- ttctgccagtgtagcaccag
TBP F – ggggagctgtgatgtgaagt	R- ttccagaaacaaaaataaggagaa
SNORD44 F – tgctgactgaacatgaaggctc	

AGO HITS-CLIP

Bead Preparations for Immunoprecipitation

Protein A Dynabeads (Invitrogen) were resuspended and washed 3x in PBS 0.02% Tween-20. 200µl beads were used per 10cm² dish of 80% confluent cells. Beads were rotated with 25µl bridging antibody (rabbit anti-mouse IgG, Jackson ImmunoResearch, 315- 005-008 at 2.3mg/ml) per 200µl beads for 35min at room temperature, then washed 3x in PBS 0.02% Tween-20. 3µl AGO antibody (2A8 ascites provided by Dr. Zissimos Mourelatos, (Nelson et al., 2007)) was loaded per 200µl beads. For irrelevant antibody controls, no 2A8 AGO antibody was added. Beads were rotated for 2hrs at 4°C then washed 3x in 1X PXL lysis buffer (1X PBS [tissue culture grade; no Mg²⁺, no Ca²⁺, GIBCO], 0.1% SDS, 0.5% NaDOC, 0.5% NP-40) containing complete protease inhibitor (mini EDTA free tablets, Roche, 1 tablet/10 ml buffer).

Cell Lysate Preparation

Media was removed from MDA-231, 4175LM2, 1833BoM2 and 831BrM2 cells, and the cells washed with cold 1X PBS before 254nm UV crosslinking 1x at 400mJ/cm² on a bed of ice (Stratalinker2400,

Stratagene). Cells were harvested by scraping in 1X PBS and pelleted. Non-crosslinked samples were collected at the same time for RNAseq analysis. Crosslinked cell lysates were prepared by adding 1ml PXL lysis buffer plus complete protease inhibitor and vortexing to disrupt cells. Lysates were DNase treated with RQ1 RNase-free DNase (Promega) for 5min, shaking at 1000rpm at 37°C. Aliquots from each cell line were pooled to the same total volume as experimental lysates and 10µl of a 1:100 RNaseA solution (diluted in lysis buffer) added for the high RNaseA control. 10µl of a 1:5000 RNaseA solution was added to all other lysates, and all samples incubated at 37°C, 5min, 1000rpm. Lysates were spun at 4°C, 30,000rpm, 20min to clear cell debris and the supernatant harvested to a new tube for immunoprecipitation (IP). Aliquots of each low RNase treated sample were again pooled for the control IgG IP.

Samples were rotated with antibody-loaded beads at 4°C for 2hs. Following IP, beads were washed sequentially (all washes 1ml) twice with 1X PXL lysis buffer, twice with 5X PXL (5X PBS [tissue culture grade; no Mg²⁺, no Ca²⁺, GIBCO], 0.1% SDS, 0.5% Na-DOC, 0.5% NP-40), twice with high stringency buffer (15mM Tris-HCl, pH7.5, 5mM EDTA, 2.5mM EGTA, 1% TritonX-100, 1% Na-deoxycholate (DOC), 0.1% SDS, 120mM NaCl, 25mM KCl), high salt buffer (15mM Tris-HCl, pH 7.5, 5mM EDTA, 2.5mM EGTA, 1% TritonX- 100, 1% Na-DOC, 0.1% SDS, 1M NaCl) and twice with low salt buffer (15mM Tris-HCl, pH7.5, 5mM EDTA) followed by two washes with PNK buffer (50mM Tris-HCl, pH7.5, 10mM MgCl₂, 0.5% NP-40). After each capture on the magnet beads, the beads were

resuspended by end-over-end rotation, with the samples washed in a random order throughout. 1.5ml SlickSeal microfuge tubes (National Diagnostics) were used throughout the CLIP protocol to reduce non-crosslinked RNA binding to tubes.

Dephosphorylation of RNA, 3' Linker Ligation, and Rephosphorylation of 5' end of RNA Tags

IPs were treated with calf intestinal phosphatase to remove the 5' phosphate from RNA crosslinked to AGO to prevent RNA circularization during ligation, an otherwise competing intramolecular reaction. Beads were resuspended in 60µl 1X dephosphorylation buffer with 3 units of CIAP (Roche) and RNasin Plus RNase Inhibitor (Promega), and incubated at 37°C, shaking at 1000rpm for 15s every 2min, 20 min. This was followed by 5X 1ml washes, once with PNK Buffer, twice with PNK buffer plus EGTA (10mM Tris pH7.5, 20mM EGTA, 0.5% NP-40), and twice again with PNK buffer (no EGTA).

For AGO CLIP, ³²P-labeled RNA linker is ligated to the 3' end of crosslinked RNA. 50pmol of puromycin-blocked linker (L32, 5'-OH-GUGUCAGUCACUCCAGCGG-3'-puromycin, Dharmacon) was labeled using T4 phosphonucleotide kinase (PNK, NEB) by adding 15µl ³²P-γ-ATP, 1µl RNasin and 2µl of T4 PNK in a 50µl total volume 1X PNK buffer and incubating for 30min at 37°C. 0.02ml 1mM ATP was added with an additional 5min incubation to drive the reaction to completion. Radiolabeled linker was purified from free ATP by spinning through a G-

25 column (Amersham). 30pmol labeled 3'RNA linker was used per 60µl T4 RNA ligase reaction (Fermentas) according to kit instructions. On-bead ligation reactions were incubated at 16°C for 1hr, 1000rpm 15s every 4min, then 80pmol of unlabeled L32 RNA linker (with 5' phosphate) added and the reaction incubated overnight. Following linker ligation, beads were washed once each with 1XPXL and 5XPXL, and 3X with PNK buffer to remove any free linker. 80µl of T4 PNK mix (6µl 1mM ATP, 4µl T4 PNK enzyme [NEB], 1µl RNasin in 1X T4 PNK buffer) was added to sample and incubated at 37°C, 1000rpm 15s every 4min, for 20 min to phosphorylate the 5' end of the AGO bound RNA for later 5' linker ligation.

SDS-PAGE Separation of RNA-binding Protein:RNA Complexes

Beads were washed three times with 1ml PNK buffer then resuspended in a 60µl solution of 1X NuPAGE loading buffer (LDS, Invitrogen), 1X NuPage Reducing Agent (Invitrogen) and 1X PNK buffer. Beads were incubated at 70°C, 1000rpm for 10min. Supernatants were run on Novex NuPAGE 8% Bis-Tris gels (Invitrogen) in SDS-MOPS running buffer (Invitrogen) for 2.5hrs, 175V, at 4°C. Two lanes per sample were used to prevent over-loading. Protein-RNA complexes were transferred to Protran BA85 nitrocellulose (Whatman) using a Novex wet transfer apparatus (Invitrogen). After transfer, the nitrocellulose was rinsed with RNase-free 1X PBS, dried, and exposed to Biomax MR film (Kodak).

Recovery of AGO-RNA Complexes, Protease Digestion, and 5' Linker Ligation

Nitrocellulose membranes were aligned carefully with the exposed film and regions of membrane excised from low RNase IP lanes at a size directly above that of the over-digested band, approximately 110-125kD for AGO-miRNA and 130-150kD for AGO-mRNA. The nitrocellulose membrane sections were cut into smaller pieces and proteinase K treated with a 0.2ml 1:5 dilution of proteinase K (4mg/ml, Roche) in PK buffer (100mM Tris-HCl pH 7.5, 50mM NaCl, 10mM EDTA)) at 37°C, 1100 rpm for 20min. 0.2ml of 1X PK buffer 7M urea solution was added and incubated for further 20min at 37°C, 1100 rpm. RNA fragments were isolated by adding 0.4ml of RNA phenol (pH 6.8, Applied Biosystems/Ambion) and 0.13ml of 49:1 CHCl₃:isoamyl alcohol, incubating a further 20min at 37°C, 1100rpm, and spinning at 20,000g in a desktop microcentrifuge at room temperature. The aqueous phase was taken to a new tube, 0.8µl glycogen (Applied Biosystems/Ambion), 50µl 3M NaOAc pH 5.2, and 1ml of 1:1 ethanol:isopropanol added, and RNA precipitated overnight at -20°C.

RNA was pelleted by spinning at 20,000g, 4°C for 30min, and washed twice with 75% ethanol. The dried pellet was dissolved in 6µl RNase-free H₂O. 1µl 10X T4 RNA ligase buffer (Fermentas), 1µl BSA (0.2 µg/µl), 1µl ATP (10mM), 0.1µl T4 RNA ligase (3U, Fermentas), and 20pmol of 5' RNA linker (RL5D) were added in a total volume of 10µl and incubated for 5hrs at 16°C with intermittent shaking. (RL5D: 5'-OH-AGGGAGGACGAUGCGGr(N)r(N)r(N)r(N)G3'-OH). Ligated RNA was

DNase digested by adding 79µl H₂O, 11µl 10X RQ1 DNase buffer, 5µl RQ1 DNase and 5µl RNasin and incubating at 37°C for 20 min, then purified by phenol:chloroform extraction (300µl H₂O, 300µl RNA phenol (Ambion), 100µl CHCl₃). RNA was precipitated from the aqueous phase overnight at -20°C by adding 50µl 3M NaOAc pH5.2, 0.5µl glycogen and 1ml 1:1 ethanol:isopropanol.

RT-PCR and High-Throughput Sequencing of PCR Products

RNA was pelleted, washed, dried as before, and then resuspended in 10µl RNase-free H₂O. 2µl of RNA per sample plus an additional 6µl H₂O were taken for a -RT control. RNA was mixed with 2µl of 5pmol/µl P32 primer (3' DNA primer 32, 5'- CTTCACCTCACCTCGCAACCG-3', Operon) and 3µl 3mM dNTPs, incubated at 65°C for 5 min in a thermocycler, chilled, and spun briefly. 1µl 0.1M DTT, 4µl 5X SuperScript RT buffer, 1µl RNasin, and 1µl SuperScript III (Invitrogen) were added and the mix incubated in a thermocycler at 50°C for 45min, 55°C for 15min, 90°C for 5min and then moved onto ice. Superscript III was omitted from -RT control reactions. PCR was performed immediately with 27µl Accuprime Pfx Supermix (Invitrogen), 0.75µl 20pmol/µl P51 (5' DNA primer 51, 5'- AGGGAGGACGATGCGG-3', Operon), 0.75µl 20pmol/ul P32 primer, and 2µl of the RT reaction, 4 reactions per sample. PCR conditions were: 95°C for 2min, 20-35 cycles 95°C 20s denature, 58°C 30s anneal, and 68°C 30s extension. Purified miRNA were amplified to 20, 22, 24 and 26 cycles, mRNA to 24, 26, 28 and 30 cycles, while -RT reactions

were taken to +4 and +6 cycles above the highest corresponding +RT reaction.

The entire PCR reaction was loaded in 95% formamide (Sigma), 5% 100mM EDTA pH8, bromophenol blue, xylene cyanol buffer in a 10% denaturing polyacrylamide gel, using 3µl Amplisize Molecular Ruler (Biorad). To visualize DNA, gels were immersed in a 1:10,000 dilution of SYBR Gold (Molecular Probes) in 1X TBE for 10min. PCR product was excised from the lowest cycle number giving visible signal, and the remaining RT reaction was then brought up to this minimal cycle number, gel purified and the products pooled. Acrylamide bands containing DNA of ~60nts were excised from miRNA samples, and ~100-150nt for mRNA. DNA was purified by crushing gel slices in diffusion buffer (0.5M ammonium acetate, 10mM Mg acetate, 1mM EDTA, 0.1% SDS), incubating at 50°C for 30min, 1200rpm, then filtering through Whatman GF/D filters in Nanosep columns (VWR). DNA was recovered from filtrate using Qiaquick gel purification buffers and columns (Qiagen) and eluted in 30µl TE. An additional PCR reaction was performed using the following fusion primers to permit sequencing on the Illumina platform.

SP5fusion: AATGATACGGCGACCACCGACTATGGATACTTAGTCAG
GGAGGACGATGC

GG3'SP3fusion: CAAGCAGAAGACGGCATACGACCGCTGGAAGTGA
CTGACAC

PCR amplification was performed using Accuprime Pfx (Invitrogen) and ranged between 6 and 10 cycles. The PCR product was separated on 2% Metaphor agarose gel (Lonza) and the lowest cycle number visible

purified using Qiaquick spin columns (Qiagen). A total of 10-30 μ l of 10nM DNA (quantified using Quant-IT, Invitrogen) was sequenced using primer SSP1 on an Illumina HiSeq.

SSP1: CTATGGATACTTAGTCAGGGAGGACGATGCGG.

All data analysis was done using the Galaxy platform (Hillman-Jackson et al., 2012) and the R Statistical Analysis Software (<http://www.R-project.org>) running edgeR (Robinson et al., 2010).

RBM47 polyA-tailed HITS-CLIP

Immunoprecipitation, CIP treatment, direct 32 P- γ -ATP labeling, SDS-PAGE and nitrocellulose transfer

Beads were prepared as per the AGO HITS-CLIP protocol, omitting the bridging antibody and using 50 μ l of ProteinG dynabeads loaded with 5 μ l anti-FLAG mouse monoclonal (Sigma F3165) per sample. Cell lysates were prepared as for AGO HITS-CLIP from duplicate dishes of UV crosslinked 1ng/ml dox treated 831BrM2 Flag-RBM47 WT, one crosslinked dox treated Tet-On control dish, and one non-crosslinked dox treated RBM47-WT dish. A lower concentration of RNaseA was used as a 1:10,000 dilution was shown to give optimal digestion of RBM47 bound fragments. IPs and subsequent washing was carried out as described. RBM47-RNA complexes were treated on bead with calf intestinal phosphatase, washed, then directly labeled on bead with 32 P- γ -ATP using 4 μ l 10X PNK Buffer (NEB), 2 μ l 32 P- γ -ATP, 2 μ l T4 PNK enzyme (NEB) and 32 μ l H₂O at 37C, 1200rpm 15s/2min, 20min. 1 μ l of non-radiolabeled 10mM ATP was added and the reaction incubated a further 5min. Beads were then washed once

with 1XPBK, once with high salt buffer, then twice with 1XPBK. SDS-PAGE and nitrocellulose transfer were carried out as for AGO with the addition of 10X Reducing Agent (NuPage) during elution of complexes from the bead.

RNA Purification, polyA Tailing and BrdU Reverse Transcription

RNA was purified as in AGO HITS-CLIP from excised membrane fragments corresponding to RBM47-WT crosslinked lanes at ~76-100kD. Pelleted RNA recovered from the first phenol:chloroform extraction were resuspended in 6.25µl H₂O and denatured at 65C for 5min with no shaking, then placed on ice. The polyA tailing reaction mix was prepared on ice by adding 1µl 10X E-PAP Buffer (Promega), 1.5µl 1mM ATP, 1.5µl RNasin Plus (diluted to 5U/µl) and 1µl E-PAP (3U/µl, NEB M0276) to the purified RNA. The reaction was incubated for 15min at 37C, then 20min at 65C in a pre-warmed thermocycler. 2.5µl of each polyA-tailed reaction was taken for a -RT control with an additional 2.5µl H₂O. Reverse transcription was carried out using 5µl polyA tailed RNA, 1µl 0.752M Tris, 1µl 8.2mM dATP, 1µl 8.2mM dCTP, 1µl 8.2mM dGTP (all Invitrogen), 1µl 8.2mM Br-dUTP (Sigma), 1µl 25µM RT Primer and 1µl H₂O. This mix was incubated for 3min at 75C then ramped down to and held at 48C. 1µl 82mM DTT, 1µl 10U/µl RNasin and 1µl SuperscriptIII (or H₂O for -RT) pre-warmed to 48C were added to the mix, then incubated for 45min 48C, 15min 55C, 5min 85C then held at 4C. RT-primers with different index sequences were used, one per replicate:

NV20T-RTAP2:

pGCACTGTTN₆GATCGTCGGACTGTAGAACTCT/idSp/CAAGCAGA
AGACGGCATAACGAT₂₀VN

NV20T-RTAP4:

pGCGAAACTN₆GATCGTCGGACTGTAGAACTCT/idSp/CAAGCAGA
AGACGGCATAACGAT₂₀VN

cDNA purification

50µl of ProteinG dynabead per sample were washed three times in 1XPBS 0.02% Tween-20, resuspended in 22µl of wash buffer plus 25µl 50X Denhardt's Solution (Sigma) and rotated at room temperature for 1hr. Beads were washed three times with 1X IP buffer (0.3X SSPE (Ambion), 1mM EDTA, 0.05% Tween-20), resuspended in 25µl IP buffer plus 25µl of anti-BrdU antibody (5g, Santa Cruz, sc-32323), rotated at room temperature a minimum of 45min then washed 3X in 1X IP buffer. The equivalent of 25µl starting volume of beads were used per anti-BrdU IP. After completion of the RT reaction, 1µl of 2U/µl RNaseH (Invitrogen) was added and the mix incubated at 37C for 20min, 10µl RNase-free H₂O added and the cDNA purified by spinning through a G-25 column. The total volume of cDNA was brought up to 40µl with H₂O, and 10µl 50X Denhardt's solution plus 50µl 2X IP buffer added, and incubated for 5min 70C, 2min 25C. This mixture was added to the equivalent of 25µl starting volume of the prepared anti-BrdU beads per sample. Tubes were rotated at room temperature for 30mins and washed once with 1X IP buffer plus 5X Denhardt's, twice with low salt buffer plus 2.5X Denhardt's, twice with

stringent buffer plus 2.5X Denhardt's, and twice with 1X IP buffer. cDNA was eluted from the beads via BrdU competitive elution by adding 50µl 100µM BrdU (Sigma) in 1X IP buffer, rotating for 30min and collecting the eluate. Eluted cDNA was purified by spinning through a G-25 column, the volume adjusted to 97.5µl with H₂O, and 37.5µl 4X IP buffer, and 15µl Denhardt's solution added. Purified cDNA was heated to 70C 5min, then 25C for 2min then added to the remaining 25µl of prepared anti-BrdU beads and rotated again for 30min, room temperature. Beads were washed once with 1X IP buffer plus 5X Denhardt's, twice with low salt buffer plus 2.5X Denhardt's, twice with stringent buffer plus 2.5X Denhardt's, and twice with 1X CircLigase Wash Buffer (33mM Tris-acetate, 66mM KCl, pH7.8).

On-bead cDNA Circularization and ApeI Linearization

cDNA was circularized by incubation for 1hr, 160C, 1300rpm 15s/30s with 2µl CircLigase 10X Reaction Buffer (Epicentre), 4µl 5M betane, 1µl 50mM MnCl₂, 0.5µl CircLigase ssDNA Ligase II (50U, Epicentre) and 12.5µl H₂O. Beads were washed by rotating for 5min twice each with low salt buffer, stringent buffer and ApeI buffer (50mM potassium acetate, 20mM Tris-acetate, 10mM magnesium acetate, pH7.9). cDNA was linearized by adding 2µl 10X NEB Reaction Buffer 4, 1.25µl ApeI (10U/µl NEB) and 16.75µl H₂O, and incubating 1hr, 37C 1300rpm 15s/30s. Beads were washed twice with low salt buffer, twice with stringent buffer and twice with Phusion wash buffer (50mM Tris, pH8.0).

PCR Amplification, DNA purification and quantification

To elute cDNA from beads 10µl 5X Phusion HF Buffer (NEB), 1µl 10mM dNTPs and 37.5µl H₂O were added, the mix transferred to a thin walled PCR tube and incubated at 98C for 45 seconds. Beads were transferred back to a microfuge tube and the eluate collected from the beads by magnet and transferred to a fresh PCR tube. To this was added 0.5µl 20µM P5 primer, 0.5µl 20µM P3 primer, 0.5µl Phusion DNA polymerase and 0.5µl 50X SYBR Green I (Invitrogen).

P5 – aatgatacggcgaccaccgacaggttcagagttctacagtccgacg

P3 - caagcagaagacggcata

PCR amplification was carried in an iQ5 real-time PCR machine in order to monitor amplification, with the samples being removed when the RFU signal reached ~1000. PCR cycle conditions were as follows: 1X 98C 30s, 15-20 cycles as necessary 98C 10s, 60C 15s, 72C 20s. PCR products were purified using MinElute columns (Qiagen) as per manufacturers instructions and quantified using Quant-It (Invitrogen). Before sequencing samples were analyzed by Tapestation (Agilent) and pooled for Illumina high-throughput sequencing using the Illumina small RNA sequencing primer – cgacaggttcagagttctacagtccgacgatc. All data analysis was done using the Galaxy platform (Hillman-Jackson et al., 2012) and the R Statistical Analysis Software (<http://www.R-project.org>) running edgeR (Robinson et al., 2010).

RNAseq Library Preparation

RNA was prepared following the TruSeq RNA sample Preparation v2 Guide (Illumina FC-122-1001). An additional size selection step was added by running cDNA on a 2% agarose gel, excising fragments at 250bp (+/- 25bp) with Genecatcher Disposable Tips (Gel Company PKB6.5) and gel purification (QIAquick Gel Extraction Kit 28704). cDNA was sequenced on an Illumina HiSeq 2000. All data analysis was done using the Galaxy platform (Hillman-Jackson et al., 2012).

3'UTR Cloning into psiCheck2 and generation of mutants

Full length 3'UTRs of human genes, or shorter fragments as indicated, were PCR amplified from MDA-MB-231 cDNA (iScript cDNA synthesis Kit, Bio-Rad) using Accuprime Supermix1 (Life Technologies) and individually cloned into psiCheck2 dual luciferase reporter vector (Promega) via XhoI and NotI digestion (New England Biolabs). Mutant constructs were made using the QuikChange Lightning Site-Directed Mutagenesis Kit with primers designed using the Agilent Technologies primer design program. Mutant 3'UTR vectors were sub-cloned into psiCheck2, and all constructs were sequence verified before transfection.

HMGA1 3'UTR F – tatctcgagcccatgcgtgccgcctgctc

HMGA1 FL 3'UTR R- tatgcggccgctccagaaaaaggatatttttttattc

HMGA1 SH 3'UTR R – tatgcggccgctccaggagggcatgtgtgtatg

HMGA1 S1 22 MUT F – ctctcactggaggaggtccttctcttctgggactg

HMGA1 S1 22 MUT R – cagycccagaaggaaggacctcctccagtgaggag

HMGA1 S1 28-5p MUT F – ccatgcgtgccgcctggaggtcactggaggagcagc
 HMGA1 S1 28-5p MUT R – gctgctcctccagtgcctccaggcggcacgcatgg
 HMGA1 S1 320 MUT F – cttccttctgggactggagaccaatgc
 HMGA1 S1 320 MUT R – ggtgggagcggagcattggtctccag
 HMGA1 S2 138 MUT F – tcaccaccacactacacagcacaggtcccgtgcagggc
 HMGA1 S2 138 MUT R – gccctgcagcgggacctgtgctgtgtagtgtggtggtga
 HMGA1 S2 142-3p MUT F- tgtgccctcaccaccactgatcacagcacaccagccgc
 HMGA1 S2 142-3p MUT R- gcggctggtgtgctgtgatcagtgggtggtgagggcaca
 HMGA1 S2 ALL MUT F - tgcctcaccaccactcttcacacct
 HMGA1 S2 ALL MUT R - ggagccctgcagcgggtcgagaggt
 CD59 3'UTR F- tatctcgagtgcagtgcagcttgagtgggg
 CD59 3'UTR R – tatgcggccgcaactcgattttcagccacttgt
 CETN3 3'UTR F – tatctcgagagaattacaaggataaacact
 CETN3 3'UTR R – tatgcggccgcttaaaattctaataaaaataaatat
 CETN3 miR335* MUT F – gcctggagccatgtgatataaaccaacttagttc
 CETN3 miR335* MUT R – gaactaagttggtttatatcacatggctccaggc
 SDAD1 3'UTR F – tatctcgagcttcttggaagttttccattc
 SDAD1 3'UTR R – tatgcggccgctatttggcattagaaacctttattg
 CDH11 3'UTR F – tatctcgagcaataacgatacaaaatttgcc
 CDH11 3'UTR R – tatgcggccgcaaaaagattgcttcttatattg
 psiCheck2 F- gtgctgaagaacgagcagtaa R – cgaggtccgaagactcattt

MicroRNA mimic transfections and dual Luciferase 3'UTR Reporter Assays

All transfections were carried out using Dharmafect Duo (Dharmacon) as per manufacturers guidelines. For Q-PCR analysis, 20×10^3 cells/well of non-reporter 4175LM2 cells were plated in 12 well format 24 hours before transfection in triplicate with 50nM specific miRNA mimic, negative control mimic A, or with transfection reagent alone (miRIDIAN mimics, Dharmacon). Cells were harvested at 48hrs. To assess the role of individual miRNAs on psiCheck2 3'UTR reporter constructs, 20×10^3 cells/well non-reporter 4175LM2 were plated in 12 well format 24 hours before transfection in quadruplicate with either 50ng or plasmid alone, or with 50nM specific miRNA mimic or negative control mimic. For endogenous miRNA regulation of 3'UTR constructs, 5×10^3 /well of non-reporter MDA-231 or 4175LM2 cells were plated 24 hours before transfection of 5ng of plasmid in 96-well format (n=6 or 10). Cells were lysed 30 hours post transfection and the ratio of Renilla to firefly luciferase measured by dual-luciferase reporter assay (Promega) using a Lumat LB9507 luminometer (Berthold Tech) as per kit instructions. Each construct was assayed in two independent experiments to verify regulation.

Generation of CETN3 overexpressing stable cell lines

The coding sequence of human CETN3 was PCR amplified from MDA-231 cDNA and cloned into pBabePuro (Morgenstern and Land, 1990) via BamHI EcoRI double digestion.

Fwd—ataggatccgccaccatgagtttagctctgagaagt

Rev—tataccggtttaatgtcaccagtcataat

Retroviruses were generated by transfecting pBabePuro-CETN3 into HEK293T packaging cells with pCMV-VSV-G & pGAG-Pol (ratio 2:1:2) using TransIT-292 transfection reagent (Mirus). Virus particles were collected 48hrs after transfection, filtered and applied to triple reporter MDA-231 and 4715LM2 miR-335OE cells in the presence of 0.8µg/ml polybrene. Virally transduced cells were selected using 2µg/ml puromycin. CETN3 overexpression was verified by Q-PCR.

Invasion assays

Cells were plated at ~50% confluence in 10cm dishes in D10F, with media changed to D0.2F (as per D10F but with 0.2% FBS) 12hrs before plating in invasion chambers. Matrigel invasion chambers (BD BioCoat™ Growth Factor Reduced BD Matrigel™ Invasion Chamber) were thawed at room temperature then equilibrated with D0.2F (500ul above and below the insert) at 37°C in 5% CO₂ for one hour. Cells were trypsinized, neutralized with D10F then washed twice with 1XPBS to remove serum before resuspending in D0.2F. The equilibration media was aspirated from the top of the invasion chamber and 500ul of 100k/ml cells added to each well. Cells were allowed to invade for 24hrs at 37°C in 5% CO₂, then membranes were washed twice with 1X PBS. Non-invaded cells were scraped from the top of the membrane, and those cells that had invaded fixed with 4% paraformaldehyde for 15min at 37°C. Membranes were washed again with 1XPBS, excised and mounted using Vectorshield

Mounting Medium plus DAPI (Vector Labs). Each cell line was tested in quadruplicate. Invasive cells were imaged using a Zeiss Axiovert S100 fluorescence microscope (4X optic) in triplicate per membrane, a cells counted using ImageJ.

Flow Cytometry Surface Staining

CD59 antibody (clone p282 (H19)) was purchased from BD Biosciences. Cell surface expression was measured using fluorescently-conjugated antibodies according to the manufacturer's instructions. Cells were resuspended in 100µl FACS buffer (PBS supplemented with 1% PHS, 1% FBS (HyClone)), 1µg of each antibody was added and incubated for 10 minutes at room temperature in the dark. Cells were washed twice with FACS buffer and collected immediately on a BD FACScaliber machine and analyzed using Flowjo software (Treestar).

Western blotting

For the test of RBM47 as an RNA binding protein, after exposure of the radioactive nitrocellulose membrane to x-ray film it was blocked for 1hr in 5% non-fat milk in PBS, followed by incubation with anti-FLAG (Sigma, F7425) overnight at 4C. The membrane was washed 4X 5min 1X TBST after each antibody incubation. Secondary anti-rabbit-HRP (Jackson ImmunoResearch Labs) was added at 1:10,000 in 5% non-fat milk 1XPBS for 40min at room temperature. HRP signal was detected by Enhanced ChemiLuminescence (Western Lightning detection kit, Perkin Elmer) using Kodak MR film. For all other western blots, cells were lysed in 1X

PXL plus protease inhibitors (mini EDTA free tablets, Roche, 1 tablet/10 ml buffer), and sonicated to shear DNA. NuPAGE loading sample buffer and reducing agent (Invitrogen) were added to 30µg protein to final concentration of 1X, and samples heated at 70C for 10min. Lysates were separated on 4-12% Bis-Tris NuPAGE gels (Invitrogen) in SDS-MOPS buffer and transferred to Protran BA85 nitrocellulose by standard methods. Membranes were blocked for 1hr at room temperature with Odyssey Blocking Buffer (LI-COR) followed by the addition of primary antibody (RBM47 Sigma HPA006347, HSP90 Cell Signaling) in LI-COR Antibody buffer (50mM Tris-HCl pH7.5, 150mM NaCl, 0.1% Tween-20, 2% BSA and 0.02% sodium azide) for two hours at room temperature or overnight at 4C. Blots were washed 4X 5min in 1X TBST after each antibody incubation. Fluorescently conjugated secondary antibody (Goat anti-mouse IgG IRDye, Goat anti-rabbit IgG IRDye LI-COR) was added in 1X TBST for 40min, blots washed 3X 1XTBST and 1X 1XTBS and fluorescence detected using an Odyssey LI-COR machine.

Alternative splicing semi-quantitative RT-PCR

Total RNA was purified from triplicate dishes of RBM47-WT#10 and RBM47 Tet-On cells treated for three days with 10ng/ml doxycycline, or MDA-231 and 831BrM2 cells as described above. Reverse transcription was carried out using random hexamers and Transcriptor Reverse Transcriptase (Roche). For each alternative splicing event investigated, pilot RT-PCR reactions were carried out with increasing cycle number to ensure amplification was within the linear range. Radiolabelled ³²P-dCTP

was added to each PCR reaction for the last two cycles. RT-PCR products were separated by electrophoresis on 6% polyacrylamide/7M Urea gels and exposed to x-ray film. In all cases, the first three lanes correspond to an equal mixture of cDNA from each cell line tested amplified at n-2, n, n+2 cycles. Lane 4 corresponds to PCR at cycle n+2 using an equal mixture of -RT control reactions. The ratio of included isoforms was determined using ImageJ analysis of scanned x-ray film. All primers were designed using the hg18 reference genome and Primer3Plus.

MDM4 F – ccctctctatgatatgctaagaaagaa	R – tagaatgtatgcatttatgctctga
MBLN1 F – catttgcaagccaagatcaa	R - aggggaacacttggtggcagat
MACF1 F – cctactcgttccagctccag	R – gcaagggatgtccgactaga
KIFAP3 F – aggagccataagtcccgatt	R - acccaaccacacagatttc
SLK F – agactatcgaacgcctggaa	R – aactctgccttctgctgctg
LIMCH1 F – gaggaataccgcaagagctg	R – cttgccatttgcgtcctct

CHAPTER 3:

DIFFERENTIAL AGO-REGULATION IN WEAKLY AND HIGHLY METASTATIC BREAST CANCER

INTRODUCTION

The metastatic progression of breast cancer has been modeled *in vivo* in mice using two general strategies, genetic engineering and transplantable model systems. The latter includes both syngenic models using mouse tumor cells, and xenograft models making use of human derived breast cancer cell lines (reviewed in Bos et al., 2010). The xenograft model chosen for this study was developed in the Massagué laboratory, using *in vivo* selection to isolate organ specific sub-lines from the metastatic human breast cancer cell line MDA-MB-231 (MDA-231) that was isolated from a patient with metastatic relapse years after removal of the primary tumor (Cailleau et al., 1978). When injected into the arterial circulation of mice, MDA-231 are able to extravasate and colonize *in vivo* the organs characteristically seen in human breast cancer patients albeit at low penetrance. By recovering those populations of cells able to metastasize to lung, bone and brain, expanding in culture and re-inoculating into mice, highly aggressively metastatic sub-lines have been developed with specific organotropisms (Figure 1.1, (Bos et al., 2009, Kang et al., 2003, Minn et al., 2005a, Minn et al., 2005b)).

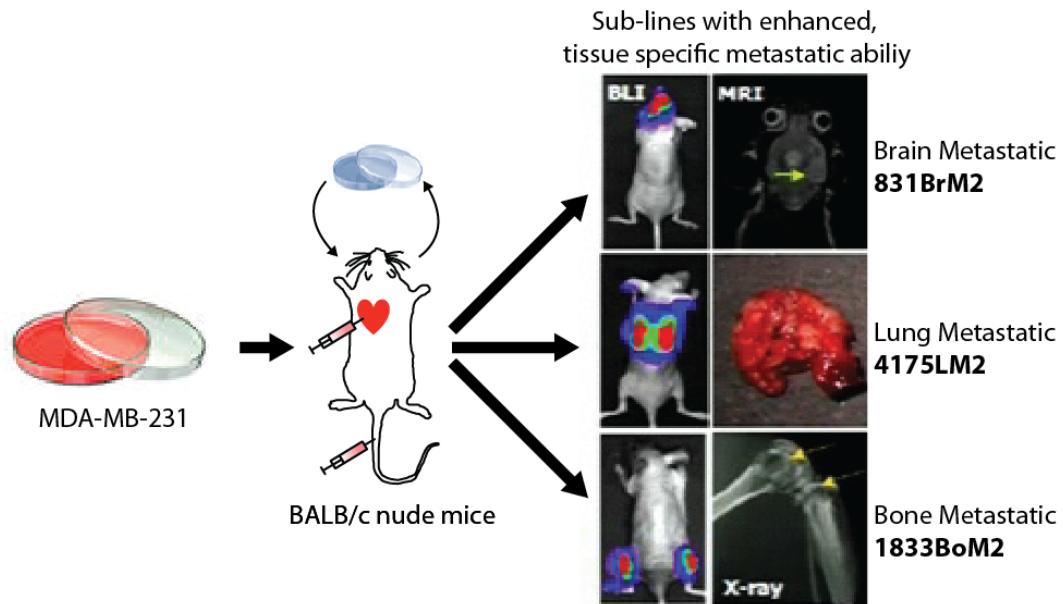


FIGURE 3.1: *In vivo* selection of human breast cancer cell lines with enhanced metastatic ability and organ specificity. “Parental” MDA-MB-231 cells, derived from the pleural effusion of a breast cancer patient with widespread metastasis years after the removal of the primary tumor (Cailleau et al., 1978), were inoculated into the arterial (intercardiac) or the venous circulation (tail vein) of immunodeficient mice allowing for systemic distribution to all organs. Organ-specific metastatic sub-lines, visualized by bioluminescence imaging of luciferase-transduced cells, were recovered, expanded in culture and re-inoculated to further enrich for highly aggressive, tissue-specific cells. The three MDA-MB-231 derivative sub-lines used in this study are highly metastatic to the bone (1833BoM2, Kang et al., 2003; Minn et al., 2005b), lung (4175LM2, Minn et al., 2005a; Minn et al., 2005b), and brain (831BrM2, Bos et al., 2009). Figure adapted from Minn et al., 2005b and Minn et al., 2005a.

Extensive molecular characterization of the aggressively metastatic sub-lines of MDA-231 has revealed that they, like their parental cell line, express a “poor-prognosis” signature for breast cancer metastasis that is associated with a clinically poor outcome and highly prognostic for the development of distal metastases (Minn et al., 2005b, van 't Veer et al., 2002, van de Vijver et al., 2002). Layered on top of this common gene set are tissue specific gene expression signatures that both distinguish the MDA-231 sub-lines from the parental line, and clinically correlate with specific site of secondary tumor formation in large cohorts of breast cancer patients. For example, those patients found to express the 54-gene lung metastasis signature identified from multiple lung metastatic sub-lines have markedly worse lung-metastasis-free survival ($p < 0.0001$) but not bone-metastasis-free survival ($p = 0.15$, (Minn et al., 2005a, Minn et al., 2007)).

More recently the regulatory role of small RNAs in the acquisition of the aggressively metastatic phenotype has been investigated using this xenograft model. Array-based profiling of miRNAs expressed in MDA-231, bone and lung specific derivative cell lines has identified miRNA expression patterns that correctly classify the three different groups, plus and additional eight miRNAs that show consistently decreased expression in all highly metastatic sub-lines (Tavazoie et al., 2008). Of these, miR-335, miR-206 and miR-126 were found to inhibit both lung and bone metastasis *in vivo* and to be markedly downregulated in patients with shorter times to metastatic relapse. Together with the previous gene expression profiles,

this data indicates the utility of this xenograft model system for identifying clinically relevant molecular events associated with tissue specific metastasis, in particular for identifying the role of AGO-miRNA mediated mRNA regulation in the suppression of metastasis.

Identification of miRNA targets has classically been carried out by searching for conserved seed matches in the 3'UTR of transcripts that show differential gene expression above an arbitrary threshold after the manipulation of a single miRNA. For example, in HeLa cells transfected with miR-124 canonical 6nt seed matches have been found in the 3'UTRs of 76% of microarray-identified downregulated transcripts, an enrichment that is reduced when transfecting with seed mutated miR-124 mimics (Lim et al., 2005). Inferring direct miR-124 regulation from this kind of data set is problematic. The bioinformatic prediction of regulatory seed sites is limited by high false positive rates and reliance on conservation (Baek et al., 2008, Selbach et al., 2008), hindering the identification of direct target gene expression levels changes that occur in a background of indirect alterations in transcript abundance. Additionally, the degree to which individual miRNAs are able to regulate targets is dependent upon the concentration of both the miRNA and target mRNA (Mukherji et al., 2011). Therefore, for any given target the degree of miRNA mediated repression may be small yet functionally relevant and may not pass the arbitrary threshold set.

While it has been proposed that at least 84% of miRNA target repression can be attributed to mRNA degradation (Guo et al., 2010) repression can occur over a broad range (Hendrickson et al., 2009, Mukherji et al., 2011) or can occur at the translational level without mRNA degradation in a target dependent manner (Eulalio et al., 2007). In certain cellular conditions, miRNA targeting can lead to activation of translation (reviewed in Vasudevan, (2012)). Given the potential for a single miRNA to coordinately regulate hundreds of targets (Hendrickson et al., 2009, Lim et al., 2005), low-level miRNA-mediated changes in multiple nodes within the same cellular pathway may sum to large physiological effects on downstream targets that are not directly regulated by the miRNA (Ebert and Sharp, 2012).

To address these limitations, we propose using HITS-CLIP to identify on a genome-wide scale direct RNA-protein interactions in living cells (Licatalosi et al., 2008), which for AGO includes both miRNA and target mRNA interactions (Figure 3.1A) (Chi et al., 2009). By mapping the fragments of mRNA that are covalently bound to AGO after UV-irradiation, we can define sites of direct endogenous interaction *in vivo*, delimiting the sequence space on a per transcript basis in which seed-match search need be carried out (Figure 3.1B). By comparing the simultaneously derived AGO-bound miRNA profile to the sequences found in robust and reproducible AGO-mRNA footprints, high confidence prediction of the miRNA (or miRNA seed family) involved in AGO targeting at any given site can be obtained (Chi et al., 2009).

Furthermore, comparative AGO HITS-CLIP can be utilized to identify sites of *de novo* AGO regulation, as was demonstrated by transfection of the neuron-specific miR-124 into HeLa cells (Figure 3.1C, (Chi et al., 2009)), or binding loss as in the genetic ablation of miR-155 in mouse T cells (Loeb et al., 2012).

By applying the AGO HITS-CLIP methodology to the MDA-231 based xenograft model of metastatic breast cancer, we aim to produce genome-wide binding maps of endogenous human AGO in both the weakly metastatic parental MDA-231 cell line and the molecularly characterized, highly metastatic bone (1833BoM2), lung (4175LM2) and brain (831BrM2) sub-lines. Given the known contribution of miRNAs to metastasis suppression in this model, we aim to identify differentially regulated targets with respect to metastatic ability and tissue tropism with no prior assumption as to the targets or resulting degree of mRNA regulation involved.

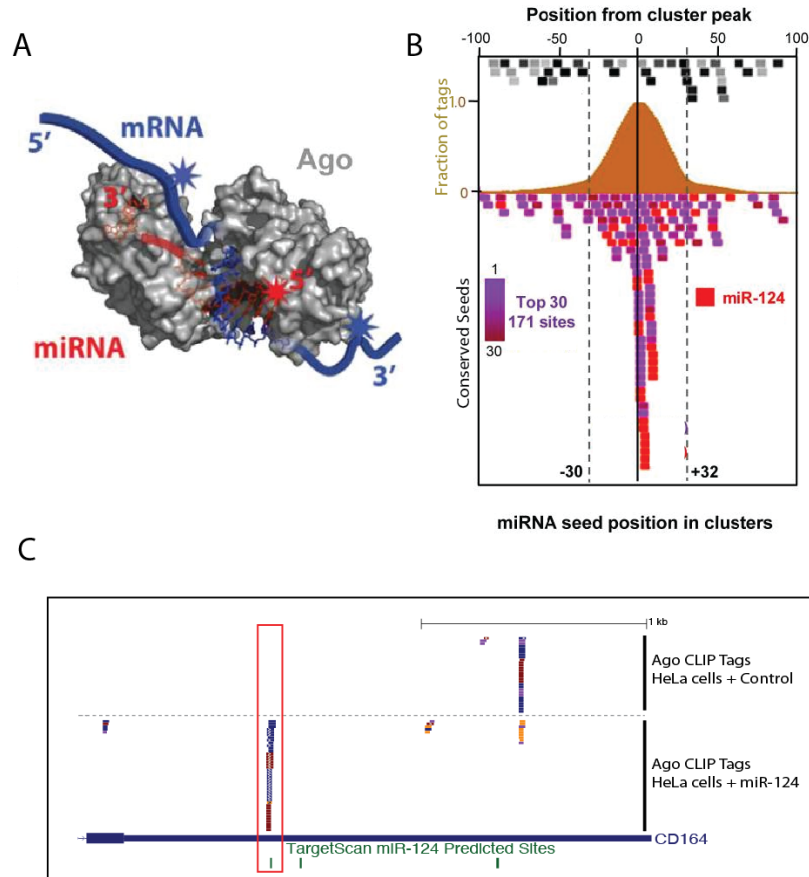


FIGURE 3.2: Argonaute HITS-CLIP produces two simultaneous data sets – AGO-miRNA and AGO-mRNA. **A:** The X-ray crystal structure of wild-type *T. thermophiles* AGO-miRNA-mRNA ternary complex (Wang et al., 2008) reveals sufficiently close contacts between AGO-miRNA and AGO-mRNA to allow for UV-induced covalent crosslink formation to both RNA species. **B:** Overlaying robust, reproducible AGO mRNA peaks (orange) reveals striking enrichment of seed sequences corresponding to the top 30 CLIPed miRNAs from the same sample (purple-red) within a 62nt AGO-bound mRNA footprint centered at the peak of AGO-mRNA binding. No enrichment of seed sequences is found for the 30 least CLIPed miRNAs (black-grey). **C:** AGO HITS-CLIP can be used to identify sites of *de novo* AGO-miRNA regulation. The top panel shows unique AGO mRNA CLIP tags from non-transfected HeLa cells on known miR-124 target CD164. The lower panel shows unique AGO CLIP tags from miR-124 transfected HeLa cells (each color indicating a biological replicate) with a *de novo* AGO-mRNA binding site centered upon a predicted miR-124 seed site (red box). (Figures 3.2A and B modified from Chi et al., 2009).

RESULTS

AGO-HITS CLIP in MDA-MB-231 and its highly aggressive metastatic derivatives with tissue tropism to the lung, bone and brain

In order to map the direct sites of AGO action on mRNAs and to identify the guide strand miRNAs in complex with AGO, HITS-CLIP was carried out in triplicate for MDA-231, 4175LM2, 1833BoM2 and 831BrM2 triple reporter expressing cell lines (herpes simplex virus 1 thymidine kinase:eGFP:luciferase), as outlined in Figure 3.3 and previously published by (Chi et al., 2009). Briefly, cells were grown in parallel until ~80% confluent, rinsed in PBS and UV-B irradiated once at 400mJ/cm². Cells were lysed, DNase treated and partially RNaseA digested before being stringently immunoprecipitated using a mouse monoclonal antibody to human AGO2 (2A8, (Nelson et al., 2007)). A portion of lysate from each sample was taken pre-IP, pooled and subsequently split for control IgG IP (to identify any non-specific pull down of RNA), and an AGO IP under high RNaseA conditions to aid in later size selection of AGO-RNA complexes.

After harsh washing, purified crosslinked AGO-RNA species were ligated to a 5'- γ P³² labeled 3'-RNA linker and size separated via SDS-PAGE. AGO-RNA complexes were transferred to a nitrocellulose membrane that was then exposed to x-ray film to allow for visualization of radiolabelled RNA (Figure 3.4A). Regions corresponding to AGO-bound miRNA and mRNA of increasing modal size were excised as indicated, proteinase K treated and the RNA purified.

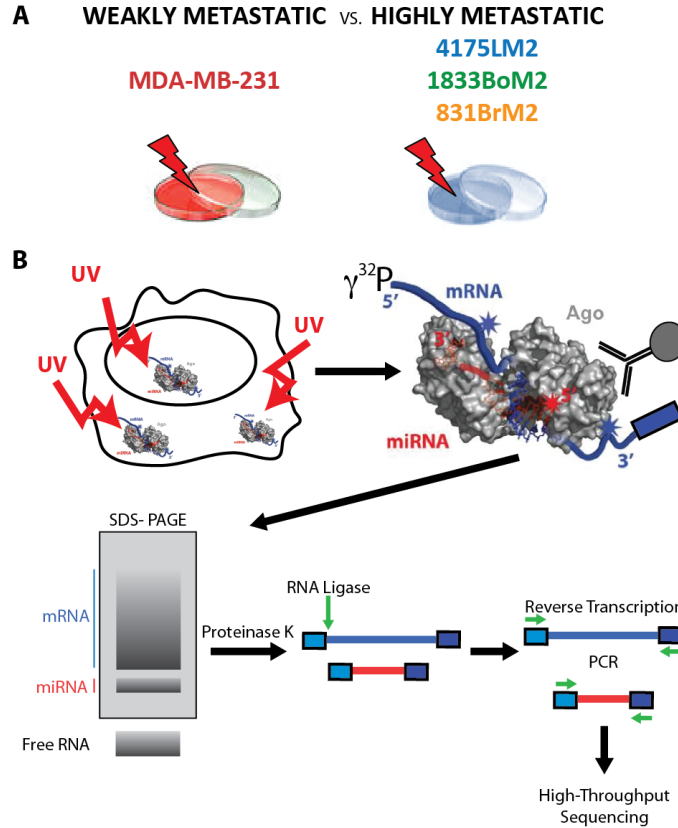


FIGURE 3.3: Argonaute HITS-CLIP schematic. **A:** 4 dishes each of weakly metastatic MDA-MB-231 cells, and the three tissue-specific aggressively metastatic sub-lines were grown to ~80% confluency. 3 dishes per cell line were UV-B crosslinked once in preparation for HITS-CLIP, with the remaining dish harvested without crosslinking for RNAseq. **B:** UV-B irradiation penetrates live cells inducing covalent bond formation between RNA and protein complexes in close association (~1Angstrom), which for AGO includes both miRNA and mRNA species. Irradiated cells are lysed and DNase treated before partial RNaseA digestion to reduce the modal size of bound RNA fragments. Purification of RNA-AGO complexes is achieved by IP under stringent conditions to remove both non-crosslinked RNA and potential contaminating proteins. While still on bead a 3'-RNA linker is ligated to the bound RNA that is then radiolabelled on the 5'end and further purified by SDS-PAGE and transfer to nitrocellulose. Bound RNA is visualized by autoradiogram to allow for size selection of sections of nitrocellulose corresponding to AGO plus ~22nt (miRNA) and ~50nt (mRNA). These sections are excised, proteinaseK treated and RNA purified. A 5' degenerate linker is ligated to allow for reverse transcription, followed by PCR amplification and high-throughput sequencing.

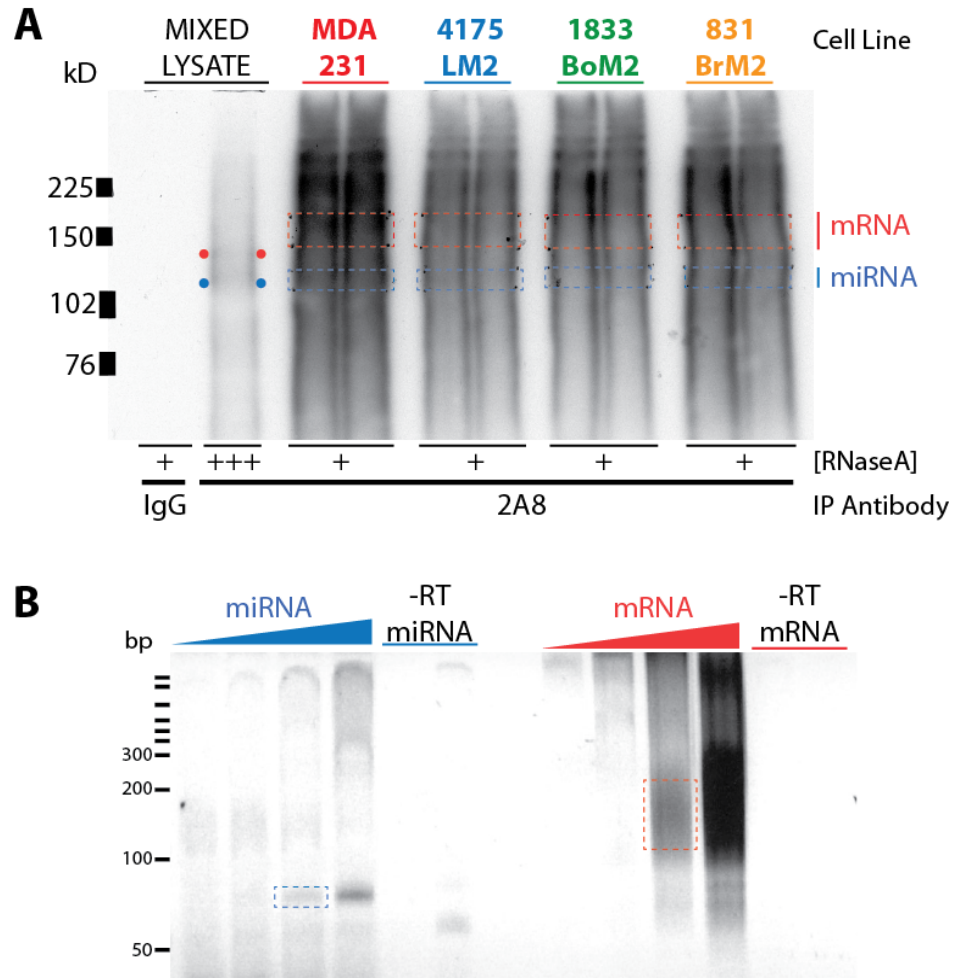


FIGURE 3.4: AGO HITS-CLIP in MDA-231 and derivative cell lines. A: Autoradiogram of crosslinked immunoprecipitates from breast cancer cell lines with irrelevant antibody (IgG lane) and 2A8, a mouse monoclonal antibody to human AGO. ^{32}P -labeled RNA is evident in only the 2A8 immunoprecipitated lysates. With high RNaseA treatment (++++) the RNA-protein complexes approach the modal size of AGO bound to miRNA (~110kD, bounded by blue dots) and AGO bound predominantly to short mRNA fragments (~130kD, red dots). With low RNaseA treatment (+), a smear of radiolabelled AGO-RNA of increasing modal size is evident for each cell line immunoprecipitation. The regions excised for purification and amplification are outlined (blue-miRNA, red-mRNA). **B:** PCR products amplified from the excised portions of the 1833BoM2 lanes in A, after 5' linker ligation and reverse transcription. cDNA from the lowest PCR cycle at which product is visible (dashed boxes) were purified and used for high-throughput sequencing. Minus RT reactions at cycle numbers +2 and +4 above that of the highest plus RT PCR reaction show no amplification.

After 5'-linker ligation and a second DNase digestion to remove any contaminating DNA, isolated RNA fragments were reverse transcribed and PCR amplified (Figure 3.4B) in preparation for high-throughput sequencing. Fragments corresponding to AGO-bound mRNA and miRNA from each replicate experiment were pooled before sequencing.

The bioinformatics pipeline used for processing AGO-bound RNAs is summarized in Figure 3.5. Reads were filtered to remove exact duplicates and low quality sequences, adapter sequences trimmed, and reads then aligned to both the human genome (hg18) and a database of mature human miRNAs downloaded from the current release of miRBase (Kozomara and Griffiths-Jones, 2011). Tags that map to single unique locations in the hg18 reference genome are collapsed to retain single representative tags and remove presumed PCR duplicate sequences based on multiple criteria; a 4nt degenerate sequence in the 5'-linker, an algorithm that estimates sequencing error through mismatches in alignment to the genome, and genomic coordinates. The multiple filtering steps applied to mRNA reads returns a highly curated set of sequences, termed "tags", representing 20-50nt mRNA fragments directly bound by AGO. miRNA reads are not collapsed in this manner due to the restriction in possible 5' degenerate linker sequences and the pre-determined length of miRNAs. The number of miRNA and unique mRNA tags obtained for each replicate AGO HITS-CLIP library are shown in Table 3.1.

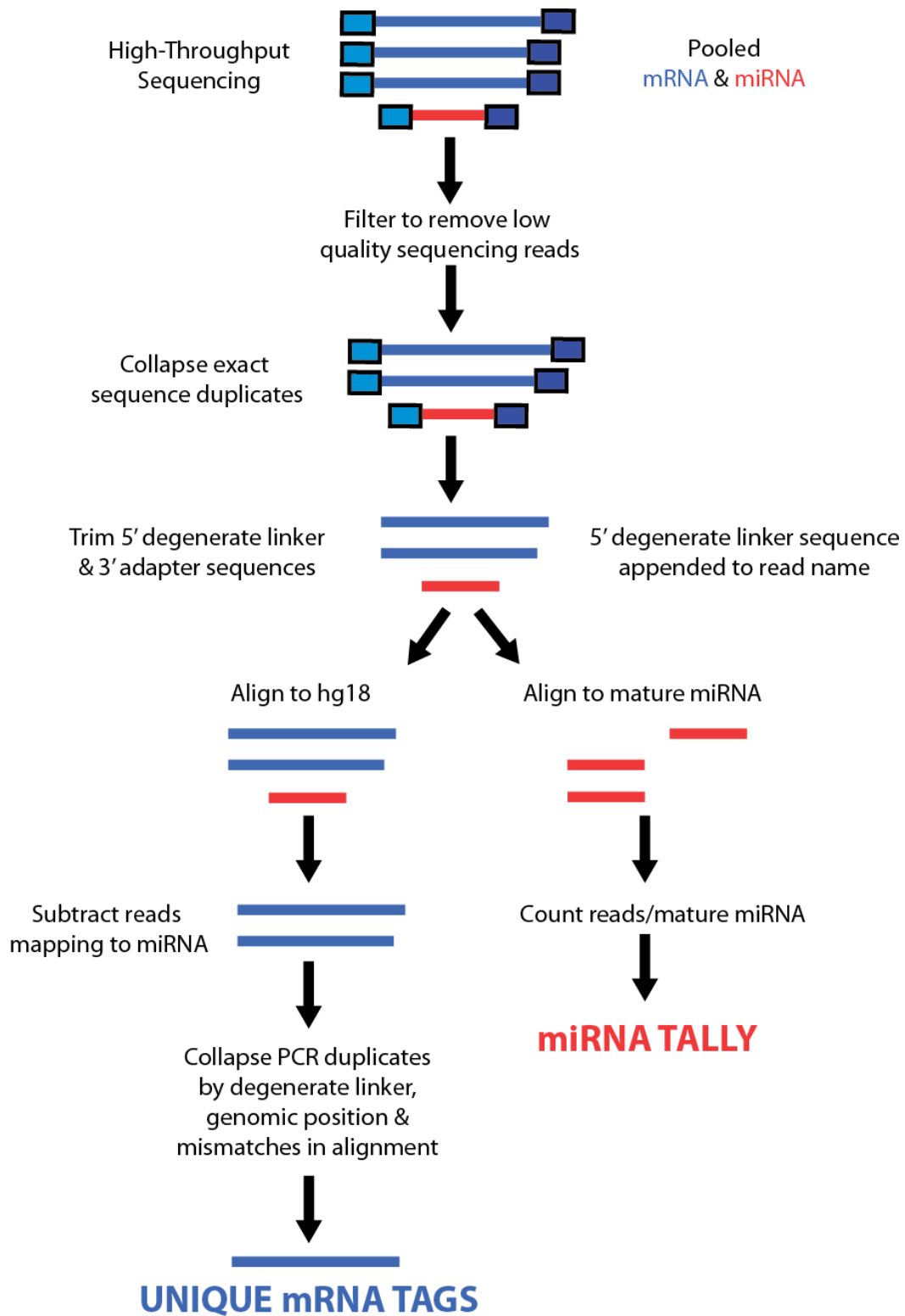


FIGURE 3.5: Bioinformatic pipeline for processing AGO HITS-CLIP high-throughput sequencing data.

TABLE 3.1: Summary of AGO HITS-CLIP sequencing results from triplicate experiments in MDA-231, 4175 LM2, 1833BoM2 and 831BrM2. Read counts are shown for each stage of data processing outlined in Figure 3.5. Total miRNA and unique mRNA tags per replicate library are bold, with the total number of unique mRNA tags for each cell line boxed. 831BrM2 sample B was sequenced on the Illumina HiSeq platform giving >10 fold higher sequencing depth than other experiments.

	MDA-231			4175LM2		
	A	B	C	A	B	C
Filtered & Adapters Removed	1,367,888	4,864,381	6,682,649	1,352,821	6,418,153	3,351,667
Mapped to hg18	810,019	2,744,822	4,031,834	908,727	3,997,760	2,027,543
miRNA Removed	694,447	1,565,276	3,192,186	747,588	2,812,665	1,179,091
miRNA Tags	182,810	1,893,836	1,238,671	240,046	1,714,987	1,178,025
Unique mRNA Tags	207,013	206,811	418,754	231,577	393,269	116,343
832,578			741,189			

	1833BoM2			831BrM2		
	A	B	C	A	B	C
Filtered & Adapters Removed	1,663,249	2,675,528	4,479,024	3,553,689	43,717,912	3,492,961
Mapped to hg18	905,789	1,725,577	2,795,921	2,142,214	28,776,799	1,970,695
miRNA Removed	903,010	1,611,590	2,055,669	2,137,285	10,216,499	1,434,244
miRNA Tags	268,248	200,310	1,065,601	433,608	24,795,582	783,884
Unique mRNA Tags	354,333	818,024	640,135	643,728	890,998	475,218
1,812,492			2,009,939			

AGO-miRNA profiles of MDA-231 and its metastatic sub-lines are highly similar

AGO-miRNA tags were mapped against a reference set of 1,100 miRNAs including star strand sequences obtained from miRBase (Kozomara and Griffiths-Jones, 2011). Individual miRNAs were deemed expressed if they were represented by a minimum of 10 tags in two or more libraries derived from an individual cell line (termed biological complexity greater than 2, or $BC \geq 2$, although it is noted that duplicate cell culture based experiments may more accurately represent technical rather than true biological replicates). 660 miRNAs were expressed representing 542 seed families, defined as groups of miRNAs sharing the same 7-mer seed sequence in positions 2-8. This number is higher than the previously reported 179 expressed miRNAs identified in MDA-231 and derivative lines by microarray (Tavazoie et al., 2008), which may be due to increased sensitivity of crosslinked AGO-miRNA deep sequencing versus array hybridization and the ongoing addition of newly identified human miRNAs to the miRBase database.

To assess the global distribution of miRNA reads between cell lines, miRNA tags per 7-mer seed family were normalized for read depth per individual HITS-CLIP library and expressed as a percentage of the total number of miRNA reads in that library, then percentages averaged across triplicates for each cell line. The top 30 miRNA seed families for each cell line ranked by average percentage are shown in descending order in Table 3.2. miR-27a/b is the most frequently cloned AGO-associated miRNA

TABLE 3.2: Top 30 miRNA seed families represented in the AGO HITS-CLIP for each cell line ranked by percentage of reads.

Seed Family	MDA-231		4175LM2		1833BoM2		831BrM2	
	%	Rank	%	Rank	%	Rank	%	Rank
hsa-miR-27a/b	29.05	1	37.77	1	24.84	1	35.12	1
hsa-miR-30a/b/c/d/e	18.23	2	16.15	2	18.86	2	18.70	2
hsa-miR-21/590-5p	9.20	3	6.93	4	7.06	3	6.32	3
hsa-miR-221/222	5.69	4	7.36	3	5.65	4	3.56	6
hsa-let-7a/b/c/d/e/f/g/i/98	3.92	5	3.99	5	4.81	5	3.08	9
hsa-miR-26a/b	3.90	6	3.86	7	3.08	9	2.28	10
hsa-miR-15a/15b/16/195/424/497	3.84	7	3.91	6	3.63	7	3.24	7
hsa-miR-181a/b/c/d/4262	3.71	8	2.86	9	3.07	10	4.27	5
hsa-miR-19a/b	3.31	9	1.31	11	3.57	8	1.97	12
hsa-miR-17/20a/20b/93/106a/106b/519d	2.92	10	2.44	10	4.62	6	3.12	8
hsa-miR-24	2.25	11	1.14	14	2.20	12	1.12	14
hsa-miR-374a/b	2.09	12	3.39	8	2.43	11	5.30	4
hsa-miR-125a-5p/b	1.57	13	1.21	13	1.68	13	2.14	11
hsa-miR-23a/23b/130a*	1.39	14	1.31	12	1.26	15	1.18	13
hsa-miR-25/32/92a/92b/363/367	0.98	15	0.49	18	1.63	14	0.72	17
hsa-miR-584	0.72	16	0.54	15	0.92	16	0.30	22
hsa-miR-29a/b/c	0.59	17	0.50	17	0.78	17	0.79	16
hsa-miR-138	0.46	18	0.18	22	0.57	19	0.25	23
hsa-miR-574-5p	0.36	19	0.05	39	0.65	18	0.07	39
hsa-miR-130a/130b/301a/301b/454/4295	0.35	20	0.20	20	0.52	20	0.33	20
hsa-miR-99a/99b/100	0.34	21	0.18	23	0.43	22	0.43	18
hsa-miR-210	0.30	22	0.45	19	0.25	27	0.11	33
hsa-miR-33a/b	0.25	23	0.07	34	0.37	24	0.11	34
hsa-miR-30a*/d*/e*	0.24	24	0.15	25	0.44	21	0.23	24
hsa-miR-22	0.20	25	0.12	27	0.30	25	0.16	26
hsa-miR-503	0.20	26	0.09	32	0.30	26	0.38	19
hsa-miR-193a-3p/b	0.19	27	0.54	16	0.20	30	0.31	21
hsa-miR-101	0.17	28	0.15	24	0.18	31	0.15	27
hsa-miR-186	0.15	29	0.11	30	0.24	28	0.17	25
hsa-miR-18a/b	0.14	30	0.05	38	0.17	32	0.12	32
hsa-miR-34a/34c-5p/449a/449b	0.13	31	0.20	21	0.10	40	0.05	42
hsa-miR-324-3p	0.13	32	0.13	26	0.23	29	0.14	29
hsa-miR-140-5p	0.13	33	0.11	29	0.09	44	0.05	40
hsa-miR-182	0.11	34	0.09	33	0.11	39	0.13	30
hsa-miR-146a/146b-5p	0.10	35	0.05	40	0.10	43	0.14	28
hsa-miR-425	0.09	37	0.11	28	0.15	34	0.09	35
hsa-miR-142-3p	0.04	48	0.01	103	0.42	23	0.79	15
Percent Total miRNA tags	97.42		98.20		95.89		97.43	

family representing between ~24% to ~38% of all miRNAs, with miR-27a being the predominant family member. The top three miRNA seed families identified by HITS-CLIP represent >50% of all the miRNA reads in each cell line. The variation in the total number of relatively rare miRNA families represented in each cell line is reflected in the total number of the most CLIPed miRNA families that together contribute 95% of the miRNA profile for each cell line (Figure 3.6). 6.14% of the total 4175LM2 miRNA profile is comprised of seed families that individually represent <1% of the total population compared to 11.13% of the 1832BoM2 profile, 8.6% in MDA-231 and 8.0% in 831BrM2.

Identification of differentially CLIPed miRNA with respect to tissue tropism

To identify potential differences in the AGO-miRNA profiles of the highly versus weakly metastatic cell lines we used EdgeR, a Bioconductor software package which examines changes in counts (AGO-miRNA reads) between two or more groups (cell lines) with replicated measurements (Robinson et al., 2010). AGO-miRNA HITS-CLIP reads for each individual miRNA were tabulated and normalized to total read depth in that library to give counts per million (cpm). miRNAs with cpm>2 in >9/12 HITS-CLIP libraries were retained for difference analysis (296 miRNAs). EdgeR was used to calculate log₂ fold change, p-values and associated FDR in a pairwise fashion between cell lines using an estimation of dispersal calculated on a per miRNA basis (“tagwise dispersal”). Given the dominance of relatively few miRNA seed families in the AGO-miRNA

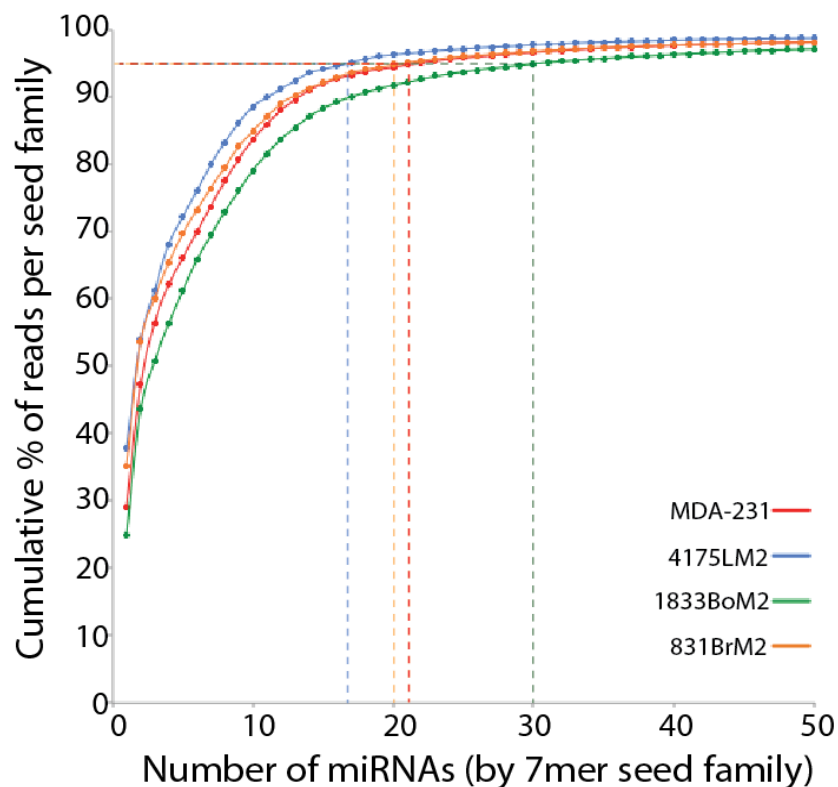


FIGURE 3.6: Distribution of miRNAs bound to AGO in MDA-231, 4175LM2, 1833BrM2 and 831BoM2 cells. The cumulative percentage of miRNA reads per seed family averaged across triplicate libraries shows that for all cell lines >95% of all reads are represented by the top 30 miRNA families. 4175LM2 cells show the least diversity in AGO-miRNA seed families with 16 7mer seed sequences representing 95% of sequenced miRNAs, followed by MDA-231 and 1833BrM (20 and 21 seed families), with 831BoM2 having the most diversity in seed families (30).

profile of each cell lines, the search for differentially CLIPed miRNAs was limited to the top 30 miRNA seed families listed in Table 3.3.

Microarray analysis of MDA-231, 4175 and 831BoM cells had previously identified 20 miRNAs specifically downregulated in bone metastatic derivatives, lung metastatic derivatives or both (Tavazoie et al., 2008). Of those 20 miRNAs previously identified, the majority of are either absent from miRNA profile (less than 10 AGO-miRNA reads, 35%) or are present at very low levels with no statistically significant difference in the AGO-miRNA profile by EdgeR analysis (60%). However, four miRNAs in the top 30 list of AGO-miRNA were found to have statistically significant miRNA tag counts with $FDR \leq 0.01$ when comparing MDA-231 to one or more metastatic sub-lines. 4175LM2 cells showed significant decrease in miR-99a, let-7c, miR-20b and miR-142-3p counts compared to all other cell lines, while both 1833BoM2 and 831BrM2 had significantly more AGO-miRNA reads corresponding to miR-142-3p than both MDA-231 and 4175LM2 (Table 3.3).

Global analysis of AGO-mRNA binding

Before looking for differential binding between cell lines, the AGO-mRNA binding events in each cell lines were characterized. Unique tags mapping to single locations on hg18 were analyzed based on two definitions of robust binding sites: clusters (Figure 3.7A) and significant peaks (Figure 3.7B) with detailed definitions of each provided in the figure legend. Briefly, “clusters” are derived from tags with minimum of 3nt

TABLE 3.3: EdgeR analysis of AGO-miRNA HITS-CLIP reads. Reads counts for each miRNA were tabulated and converted to counts per million (cpm) on a per library basis to normalize for read depth. miRNAs with cpm>2 in >9 of the 12 libraries were retained for further analysis (296 miRNAs). Using EdgeR, log2 fold change in miRNA count, p-values and FDR were calculated in a pairwise fashion between cell lines using an estimation of dispersal calculated on a per miRNA basis (“tagwise” dispersal, see (Robinson et al., 2010)). miRNAs in top 30 seed families with significant difference in CLIP tag counts between more than two cell lines are shown. Consistent significant decrease in the number of CLIP AGO-miRNA reads for miR-142-3p, miR-99a, miR-20b and let-7c is seen in 4175LM2 cells when compared to all other cells, while a significant increase in miR-142-3p is seen in both 1833BrM2 and 831BoM2. Line indicates no significant difference as calculated by EdgeR.

miR-142-3p	Log2 FC	p-value	FDR	miR-99a	Log2 FC	p-value	FDR
LM2/MDA	-2.64	1.10E-05	0.0019	LM2/MDA	-6.23	1.83E-05	0.0019
BoM/MDA	3.99	1.27E-07	3.76E-05	BoM/MDA	-	-	-
BrM/MDA	4.67	1.05E-10	3.10E-08	BrM/MDA	-	-	-
BoM/LM2	6.63	3.74E-17	1.11E-14	BoM/LM2	6.91	3.82E-07	5.65E-05
BrM/LM2	7.31	1.05E-20	3.10E-18	BrM/LM2	4.65	2.87E-07	4.25E-05
BrM/BoM	-	-	-	BrM/BoM	-	-	-

let-7c	Log2 FC	p-value	FDR	miR-20b	Log2 FC	p-value	FDR
LM2/MDA	-3.30	4.37E-05	0.0032	LM2/MDA	-3.64	0.0007	0.0223
BoM/MDA	-	-	-	BoM/MDA	-	-	-
BrM/MDA	-	-	-	BrM/MDA	-	-	-
BoM/LM2	3.83	4.88E-06	0.0005	BoM/LM2	3.54	0.0003	0.0099
BrM/LM2	2.53	5.00E-06	0.0004	BrM/LM2	3.55	2.91E-05	0.0014
BrM/BoM	-	-	-	BrM/BoM	-	-	-

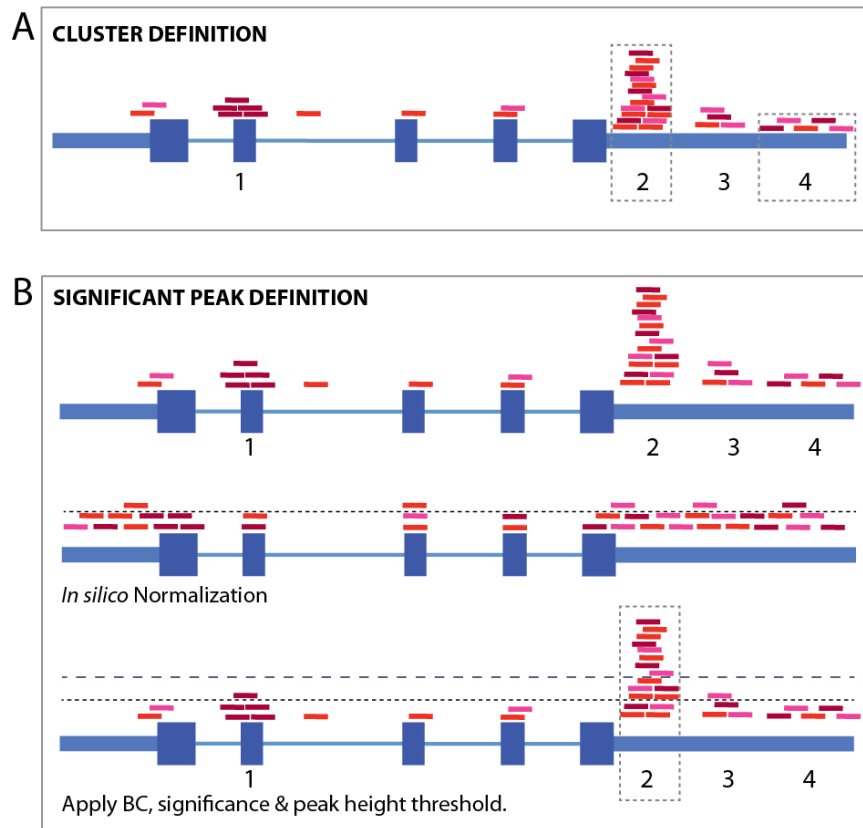


FIGURE 3.7: Definition of HITS-CLIP clusters and significant peaks **A:** Schematic representation of cluster definition. Genomic localization of unique mRNA tags from triplicate experiments from one cell line (red lines, each shade representing a replicate library) are shown on a pre-mRNA transcript (blue). Clusters are considered reproducible if they contain tags derived from ≥ 2 replicates that overlap by a minimum of 3 nucleotides (biological complexity, $BC \geq 2$), and robust if they contain a minimum scaled number of unique tags proportional to the total read depth of the cell line ($\text{tag} \geq 5$ - MDA & LM2, ≥ 12 - BoM2, ≥ 13 - BrM2). Therefore, binding sites 2 & 4 would be considered clusters in MDA or LM2 despite differing peak height and widths, but only site 2 would be a cluster in BoM2 or BrM2. **B:** Schematic representation of significant peak definitions with binding events as described in A. *In silico* normalization on a per transcript basis defines a threshold of significance (dotted line) for AGO-binding sites based on the observed vs. expected peak height calculated from random shuffling of unique tags along the exonic regions of the gene. Reproducibility and robustness were determined as for clusters (dashed line) therefore only binding site 2 is considered a significant peak.

overlap that are reproducible ($BC \geq 2$), and contain a minimum scaled number of unique tags proportional to the total read depth of the cell line. “Significant peaks” represent those clusters whose maximal number of overlapping tags (peak height) is above that of the expected peak height resulting from random shuffling of all CLIP tags mapping to that transcript along its genic length. As per (Chi et al., 2009), a narrow footprint of ± 32 nt centered on the significant peak is used to determine the AGO footprint while clusters are delimited by the genomic coordinates of the most exterior tags in that cluster. Significant peak analysis allows for more stringency in defining binding sites by narrowing the footprint of AGO binding to -30 nt/ $+32$ nt centered on the peak which is useful for identification of the regulatory miRNA involved in targeting (Chi et al., 2009). However, cluster analysis is more sensitive when searching for differential binding, allowing for identification smaller clusters may not reach statistical significance due to a background of robust, common binding elsewhere on a transcript.

The threshold for the minimum number of tags required to define a binding site (cluster or significant peak) was set at 5 tags for MDA-231 and 4175LM2 cells, with the threshold increasing proportionally to 12 and 13 tags for 1833BoM2 and 831BrM, respectively due to ~ 2.5 fold greater unique tag numbers in these cell lines. All binding sites were required to be $BC \geq 2$, and those mapping to chrY, RNA genes (defined by the RNAgenes track on the UCSC genome browser) and pre-miRNA loci were

removed from the analysis. Setting these thresholds defined 6947, 8802, 7716 and 8514 clusters, and 3422, 2636, 3233, and 3259 significant peaks in MDA-231, 4175LM2, 1833BoM2, and 831BrM cells respectively.

The genomic localization of $BC \geq 2$ clusters was highly reproducible between cell lines, with the majority of binding occurring in genic regions (5'UTR, coding sequence and 3'UTR, Figure 3.8A). Despite the well-defined role of Argonaute in regulation of transcripts via interactions in the 3'UTR and to a lesser extent in coding sequence and 5'UTRs, a surprisingly high proportion of binding was evident in intronic regions. While the biologic consequence of AGO intronic binding is not fully understood, it is highly reproducible and has been consistently observed in published AGO HITS-CLIP data sets (Chi et al., 2009), (Loeb et al., 2012), including those derived from Dicer null mouse embryonic stem cells (Leung et al., 2011), and multiple unpublished data sets (Darnell Lab). Annotation of significant peaks revealed a relative enrichment of 3'UTR binding (Figure 3.8B).

AGO-mRNA binding patterns are highly reproducible between cell lines

Unique tags from all 12 AGO HITS-CLIP mRNA libraries were pooled and clustered as one data set to define all possible AGO binding sites. Cluster analysis was chosen rather than significant peak analysis due to the uneven sequencing depth between cell lines. Clusters were filtered to remove those that did not meet the minimum criteria outlined in Figure 3.7A in any one cell line ($BC \geq 2$ tags ≥ 5 in MDA & LM2, tags ≥ 12

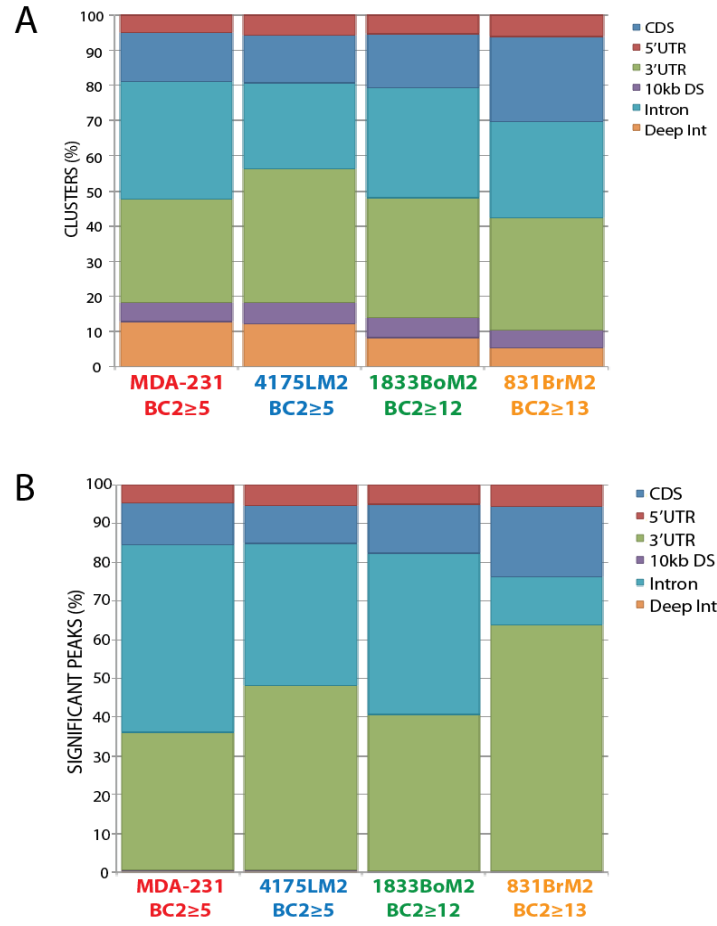


FIGURE 3.8: Genomic distribution of reproducible AGO-mRNA binding. **A:** Unique mRNA reads from each cell line were clustered based on genomic coordinates and annotated to give genomic location. Binding sites on chrY, RNA genes (coordinates derived from the RNAsenes track from UCSC genome browser) and pre-miRNA loci were removed before annotation. **B:** Reproducible significant peaks as defined in Figure 3.7B were annotated as in A.

BoM, or tags \geq 13 BrM), leaving a total of 18,444 robust and reproducible AGO binding sites. To globally assess AGO binding between the four cell lines, the number of tags contributed by each cell line in each cluster were plotted and pairwise correlation coefficients calculated (Figure 3.9). The global distribution of AGO-mRNA tags in clusters from each of the four cell lines were highly reproducible, with correlation coefficients between 0.80 and 0.92, indicating that the majority of AGO binding is common to all cell lines. The top two clusters, corresponding to the full-length transcripts of the non-coding RNAs MALAT1 and NEAT1, were removed from this analysis as inclusion artificially increased the correlation coefficients.

Bioinformatic identification of differential AGO binding sites

To identify AGO-mRNA binding sites disproportionally enriched in one of two compared cell lines, the total number of tags per cluster from each cell line were analyzed in a pair-wise fashion, normalized to the total number of tags in clusters in that cell line. Fold change and associated p-value for each cluster were calculated using scripts developed in-house by Chaolin Zhang for the Galaxy bioinformatics platform (Hillman-Jackson et al., 2012), and an FDR derived using the Benjamini-Hochberg correction for multiple hypothesis testing. Clusters were annotated to give both gene identity and genomic location, with differential binding search subsequently limited to 3'UTRs.

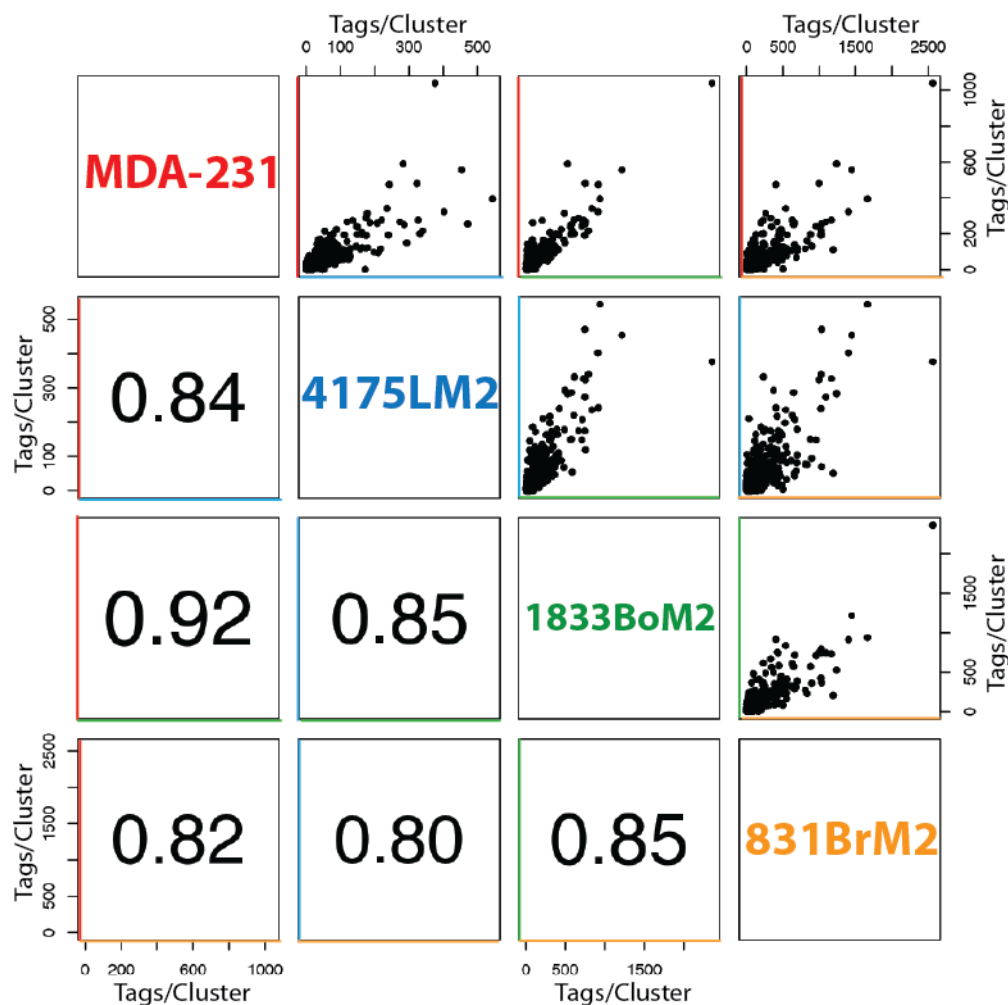
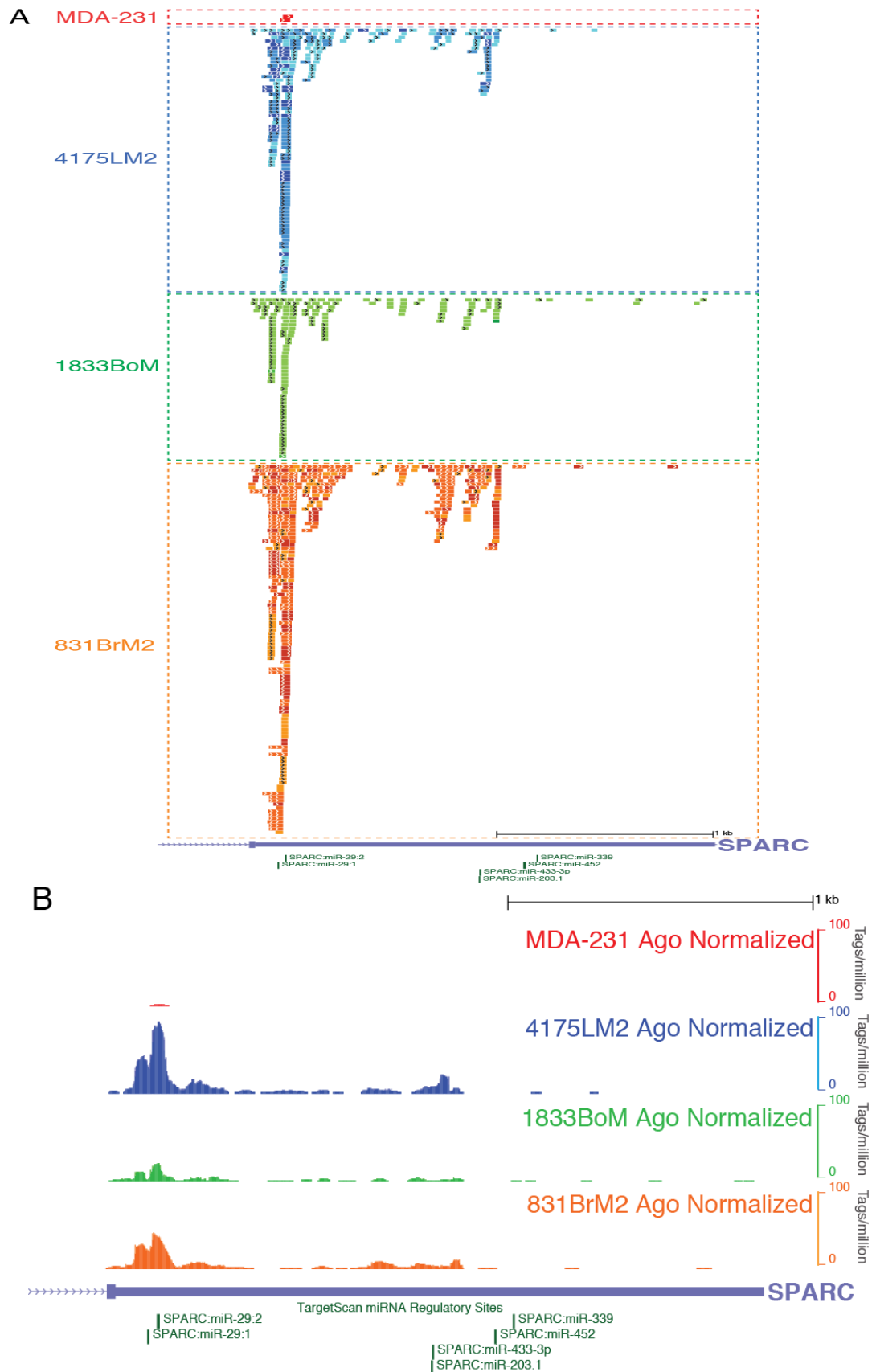


FIGURE 3.9: Pairwise correlation of tags in AGO-mRNA clusters. All unique tags from the 12 HITS-CLIP experiments were pooled and clustered with minimum 3nucleotide overlap. Clusters meeting the minimum definition of a robust and reproducible cluster in any single cell line were retained (BC2 tags \geq 5 - MDA or LM2, BC2 \geq 12 - BoM2, or BC2 \geq 13 - BrM2, 18,466 clusters), and the number of tags contributed from each cell line per cluster was counted. Upper half of figure: Pairwise scatter plots of unique tags per cluster for each cell line as indicated. Lower half of figure: absolute value of the Pearson correlation coefficient of tags per cluster.

The most robust differential binding site commonly gained in the three metastatic sub lines compared to MDA-231 was located in the 3'UTR of secreted protein, acidic, cysteine-rich (osteonectin), SPARC. The 3'UTR of SPARC is shown in Figure 3.10A, with the unique mRNA tags from each library shown as horizontal bars grouped by cell type (indicated by color), with each individual replicate library shown in a different shade. Multiple peaks of BC=3 AGO binding are evident in the 4175LM2 and 1833BrM data sets, while AGO binding is relatively absent in the parental MDA-231 cells. The 831BoM2 data set predominantly contains tags from one replicate HITS-CLIP library (118/120 tags), therefore after visual inspection this cluster was not deemed reproducible despite being BC=2. Given the variable sequencing depth between cell lines, the number of overlapping unique tags at any given genomic position were calculated and normalized to the total read depth in that cell line (tags per million unique tags) in order to more easily visualize binding differences (Figure 3.10B). The major AGO-binding site on SPARC is centered on two previously validated, TargetScan predicted miR-29 seed sequences (Kapinas et al., 2009, Wang et al., 2012).

SPARC has previously shown to be highly expressed in 4175LM2 and 831BrM2 cells, but is absent from MDA-231 and bone metastatic sub-lines (Minn et al., 2005a). Therefore the *de novo* AGO binding identified in 4175LM2 and 831BrM2 when compared to MDA-231 occurs as a function of target expression level. While this type of binding is not informative for the identification of alterations in AGO-mediated transcript regulation

FIGURE 3.10: Identification of *de novo* AGO binding sites on the SPARC 3'UTR in metastatic derivatives of MDA. The most robust changes in AGO-mRNA binding between cell lines identified by difference analysis occurs in the 3' UTR of SPARC, a gene previously shown to be highly expressed in 4175LM2 and 831BrM but absent from MDA-231 & bone metastatic sub-lines. **A:** Unique tags are plotted for each cell line as indicated by color, with different hue indicating tags derived from different replicates. Several robust, reproducible AGO-binding sites are evident on the 3'UTR of SPARC in LM2 and BoM2 cell lines, with the largest peak proximal to the stop codon and centered on two miR-29 seed sites. Although BC2 binding is evident in BoM2 cells, these clusters are predominantly comprised of unique tags from one data set (118/120 tags from BoM2 replicate A) and therefore are not reproducible. MDA cells lack any robust, reproducible AGO binding sites, consistent with the lack of expression of SPARC in these cells. **B:** The same view of SPARC as in A, showing normalized AGO-mRNA binding calculated on a per cell line basis as: number of overlapping unique tags/(total number of unique tags/ 10^6). Normalization reveals that the poorly reproducible BoM2 AGO binding occurs at a much lower level than that in LM2 and BrM2 consistent with the low level of expression of SPARC in BoM cells.



with respect to gain in metastatic potential, it is a valuable independent read out of transcript steady state mRNA level changes between cell lines. Inspection of all genes previously identified as part of the 54 gene lung metastasis signature revealed *de novo* AGO binding events in the majority of transcripts known to be upregulated in 4175LM2 cells relative to MDA-231, while known downregulated genes showed no evidence of additional AGO regulation in 4175LM2 cells (data not shown). This was not unanticipated given that the relative impact of miRNAs on a target transcript levels is generally small (Ebert and Sharp, 2012), and the expression level changes of the genes defining the lung metastasis signature were high, 3-fold or greater, although once expressed the majority show evidence of robust AGO-mRNA binding.

To control for transcript steady state levels between cell lines, previously published Affymetrix microarray data was incorporated into the difference analysis (Kang et al., 2003, Minn et al., 2005a, Bos et al., 2009), with the assumption that loss of AGO binding, and hence loss of negative regulation, should lead to increase in expression and vice versa. Given that non-expressed targets cannot by definition be bound by AGO, differentially bound transcripts were required to be expressed in both cell lines in which differential binding was detected. RNAseq libraries were prepared from non-crosslinked dishes of cell harvested at the same time as those samples used for AGO HITS-CLIP analysis, however due to low numbers of reads obtained from 1833BoM2 and 831BrM2 samples these were not reliable indicators of differentially expressed transcripts across

the four cell lines. The RNAseq profiles for MDA and LM2 were used to corroborate the previously published microarray profiles (data not shown).

Identification of AGO-mRNA binding sites gained in metastatic derivatives of MDA-231

Two 3'UTR clusters were found to have significantly enriched AGO binding in all three metastatic sub-lines ($p \leq 0.05$) that was coupled with reduced expression levels when compared to MDA-231. These clusters were located in an ~1kb region of cyclin-dependent kinase 6 (CDK6) spanning the terminal exon and extending into the 3'UTR, and in an ~330bp region of the tumor suppressor candidate 2 (TUSC2/FUS1). The AGO binding sites on the 3'UTRs of CDK6 and TUSC2 are shown in Figure 3.11, with the AGO clusters identified as being significantly differentially bound highlighted. Both clusters identified are comprised of multiple peaks of AGO binding embedded in a background of low-level scattered tags. Adding to the complexity of regulation is a background of extensive AGO binding in other regions of the 3'UTR that is particularly pronounced in CDK6, with multiple cell type specific gains and losses in binding that may play equally as important a role in down regulation of the transcripts in metastatic sub lines relative to the parental MDA-231 cells.

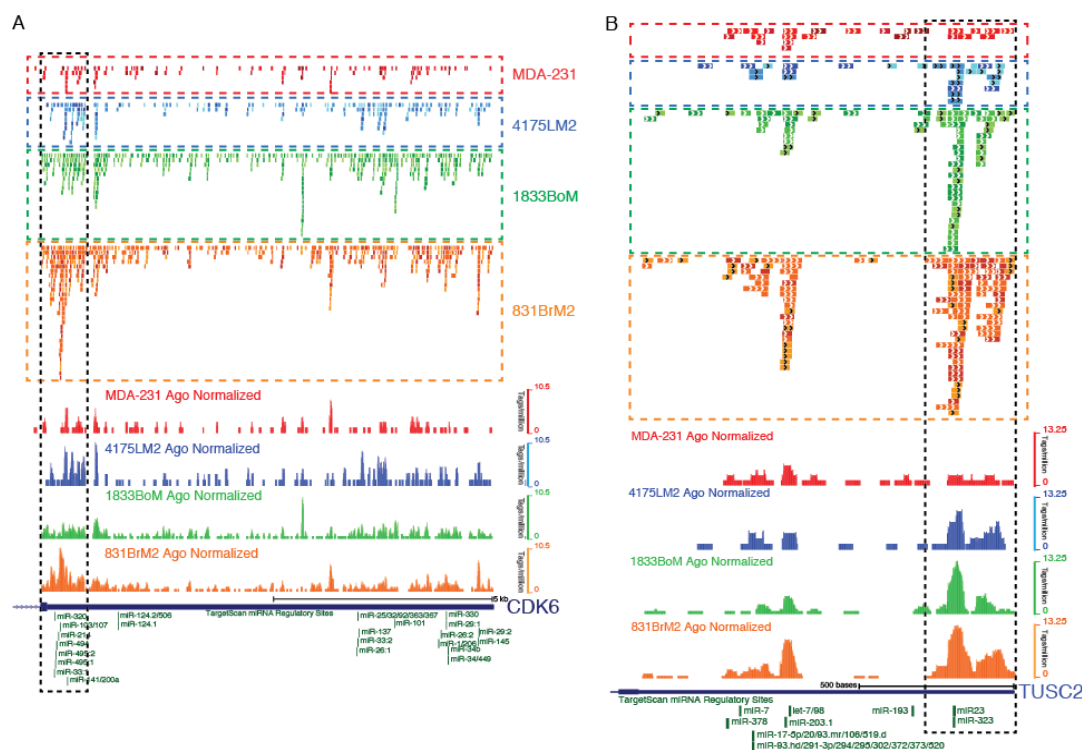


FIGURE 3.11: Identification of AGO-mRNA binding sites gained in metastatic derivatives of MDA-231. A: CDK6 shows gain of AGO binding a section of the 3'UTR spanning in an ~1kb region containing multiple individual peaks of binding (black dashed box). Upper panel: Unique mRNA AGO tags aligned to the 3'UTR of CDK6. Lower panel: Normalized binding maps reveal increased AGO binding at the proximal end of the annotated 3'UTR in the three metastatic sub-lines relative to the parental MDA-231. The number of tags in the highlighted cluster is increased 2.5-fold in 4175LM2 ($p=0.0003$, FDR 0.039), and 1.7-fold in 1833BoM2 cells ($p=0.031$, FDR=0.316) and 2.5-fold in 831BrM2 cells ($p=2.56e^{-5}$, FDR= 0.027). **B:** TUSC2 3'UTR contains a cluster gained in all three metastatic sub lines when compared to MDA-231. The number of tags in the highlighted cluster is increased 2.5-fold in 4175LM2 ($p=0.022$, FDR 0.498), and 2.4-fold in 1833BoM2 cells ($p=0.011$, FDR=0.215) and 2.8-fold in 831BrM2 cells ($p=0.002$, FDR= 0.019).

Identification of AGO-binding sites commonly lost in highly metastatic sub-lines

To identify loss of 3'UTR AGO-binding events common to all highly metastatic cell line derivatives of MDA-231, clusters were filtered to return positive and significant fold changes in tags per AGO CLIP clusters (MDA/sub-line, $p \leq 0.01$) with corresponding transcript level changes greater than 0.8 fold by microarray. Unexpectedly, only three clusters passed this threshold (Table 3.4), located in ras-related C3 botulinum toxin substrate 1 (RAC1), high mobility group AT-hook 1 (HMGA1) and tumor protein p53 binding protein, 2 (TP53BP2). Subsequent Q-PCR analysis of TP53BP2 levels in MDA-231 and LM2 cells revealed strong down regulation in LM2 cells contrary to the microarray data (6.4 fold decrease in LM2, $p = 0.0017$, data not shown). HMGA1 was selected for follow up assays due to the relatively high tag counts in AGO clusters, positive confirmation by Q-PCR of increased steady state transcript levels both 4175LM2 and 831BoM2 cell lines correlating with loss of AGO-mRNA binding (Figure 3.12), and the extensive literature associating increased HMGA1 expression with multiple aggressively metastatic cancers (Fedele and Fusco, 2010).

Differential AGO binding in the 3'UTR of HMGA1

The HMGA1 3'UTR cluster identified by differential AGO binding analysis is proximal to the stop codon (Figure 3.13, black dashed box). Multiple additional peaks of AGO binding with equal normalized peak height are seen in all four cell lines at more distal 3'UTR positions, suggesting specific loss of AGO binding only in the identified cluster.

TABLE 3.4: AGO 3'UTR binding sites lost in all three highly metastatic sub-lines of MDA-MB-231. FC= fold change, FDR = false discovery rate (Benjamini-Hochberg correction).

GENE	Ago- mRNA Unique Tags									Expression				
	MDA-231	4175LM2	1833BoM	831BrM2	MDA-231/4175LM2		MDA-231/1833BoM		MDA-231/831BrM2		Affy ID	LM2 MDA	BoM MDA	BrM2 MDA
					FC	p-val (FDR)	FC	p-val (FDR)	FC	p-val (FDR)				
RAC1	12	0	3	0	10.89	0.005 (0.262)	7.51	1.12e-4 (0.011)	34.21	5.46e-7 (2.91e-5)	208640_at	1.30	1.01	1.06
TP53BP2	27	2	8	10	7.82	3.21e-5 (0.008)	7.19	1.56e-8 (6.69e-6)	6.70	1.17e-8 (9.66e-7)	208641_s_at	1.19	1.19	1.78
											203120_at	1.66	1.25	1.98
HMGA1	27	5	24	36	3.910	8.74e-4 (0.090)	2.59	7.82e-4 (0.044)	1.99	7.76e-3 (0.045)	206074_s_at	1.54	1.26	0.88
											210457_x_at	1.39	1.45	0.53

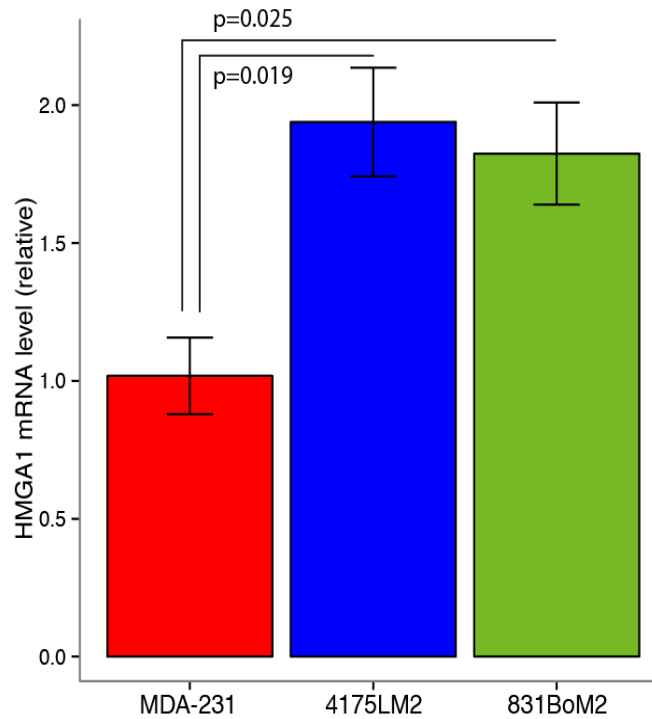


FIGURE 3.12: Loss of AGO binding correlates with increased steady state levels of the HMGA1 transcript. HMGA1 mRNA steady state levels are higher in 4175LM2 and 1833BoM2 cells relative to the parental MDA-231. Normalized to TATA-binding protein (TBP), selected as the least variable housekeeping gene across the four cell lines from a panel of ten potential normalization genes (GeNorm analysis (Vandesompele et al., 2002)). n=3, error bars = s.e.m, p-values calculated by t-test.

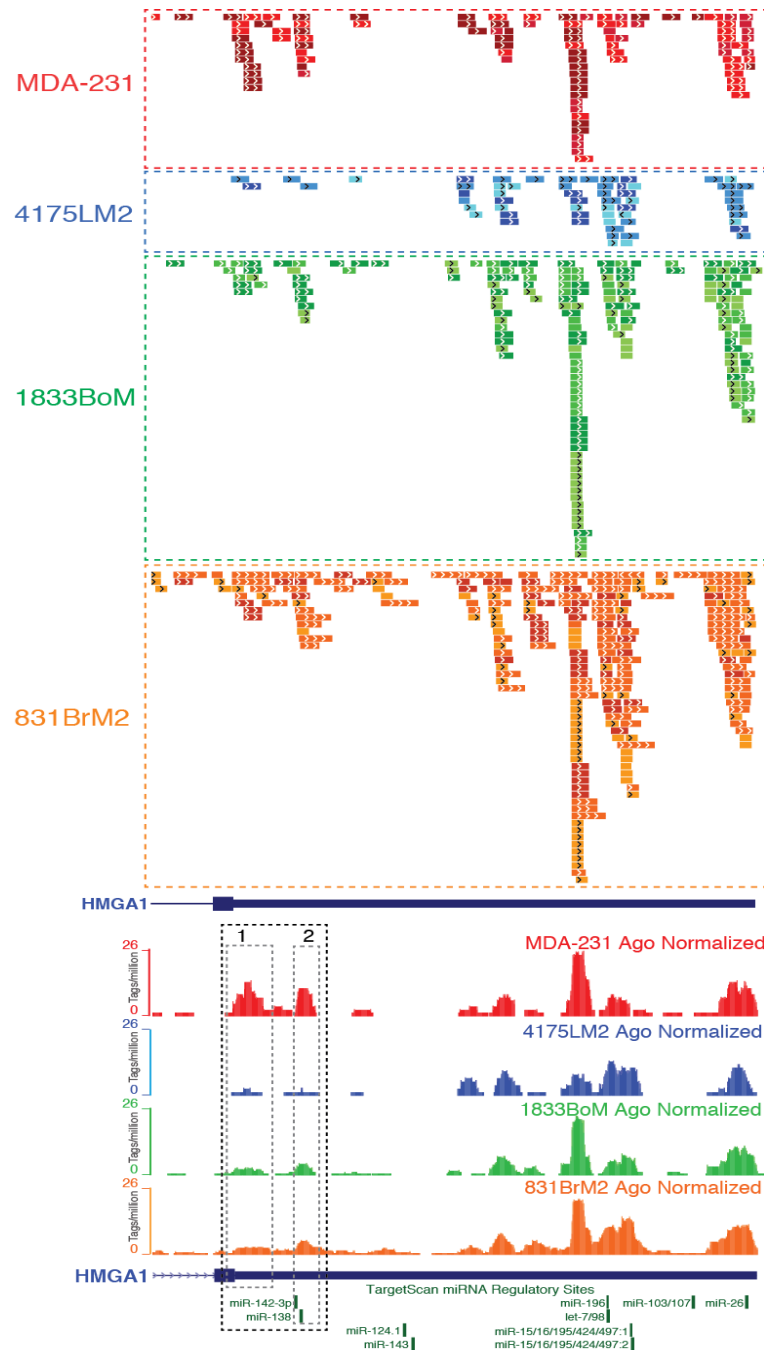


FIGURE 3.13: HMGA1 shows evidence of differential AGO binding between parental and metastatic derivative sub-lines. Upper panel: Unique mRNA AGO tags aligned to the 3'UTR of HMGA1. Lower panel: Normalized binding maps reveal decreased AGO binding X proximal to the stop codon in the cluster identified in Table. Binding on other regions of the transcript is unchanged suggesting specific loss of AGO regulation at this site. The two peaks of binding within the differentially bound cluster are labeled as site 1 and 2.

No alternative splicing of the HMGA1 3'UTR in the region bound by AGO differential binding was evident in EST tracks on the UCSC genome browser, the RNAseq profiles of MDA-231 and 4175LM2, or by RT-PCR analysis (data not shown). The MDA-231 enriched cluster is comprised of two individual peaks of binding, designated site 1 and 2, with low levels of adjoining tags bridging the two peaks. None of the published HMGA1 miRNA regulatory sites are located within the differentially bound cluster region of the HMGA1 3'UTR (D'Angelo et al., 2012, Kaddar et al., 2009, Palmieri et al., 2012, Wei et al., 2011). Shown below the HMGA1 3'UTR in Figure 3.13 are Targetscan miRNA target predictions showing miR-142-3p and miR-138 regulatory sites overlapping binding site 2, while no miRNA seed sites are predicted for site 1. miR-142-3p was significantly lower in the 4175LM2 AGO-miRNA profile than MDA-231 by EdgeR consistent with a loss of binding, however, however miR-142-3p levels are significantly increased in both 1833BoM2 and 831BrM2 inconsistent with loss of binding in these cell lines (Table 3.3). miR-138 is not within the top 30 most CLIPed miRNA in any cell line, but shows slight decreases in AGO-miRNA profile in both 4175LM2 (log2 FC=-1.5, p=0.03, FDR=0.35) and 831BoM2 (log2 FC=-2.0, p=0.14, FDR=0.99) when compared to MDA-231. The miRNA seed matches in binding site 1 corresponding to the most frequently cloned AGO-miRNA reads include miR-22, a top 30 miRNA, and miR-320a/b/c. Neither of these miRNAs showed significant differences in AGO-miRNA profile by EdgeR analysis.



FIGURE 3.14: Seed sites of expressed miRNAs in the differentially bound region of the HMGA1 3'UTR. The sequences defined by the two peaks of MDA AGO-mRNA binding were searched for seed matches to miRNAs expressed in MDA-231 cells as determined by the AGO-miRNA profile. Site 1 contains a seed match to a top 30, miR-22 (see Table 3.2), plus seed matches to miR-320a (ranked 45th MDA, 50th LM2, 35th BoM, 48th BrM). Mutant constructs were made as indicated in either in the context of full-length HMGA 3'UTR or in a shorter fragment corresponding to chr6: 34,320,641-34,320,915. miR-28-5p mutants were used as a negative control due the consistent, low level representation in the miRNA profile of all four cell lines (<0.015%). Site 2 contains seed matches to both miR-142-3p and miR-138.

To determine whether the loss of AGO binding to HMGA1 in metastatic cells occurs as a result of any of the above AGO-miRNA interactions, dual luciferase reporter constructs were made in psiCheck2 using either full length HMGA1 3'UTR or a shorter fragment corresponding to the region spanning the differential AGO-binding (chr6: 34,320,641-34,320,915). The rationale behind using a smaller fragment was based on the hypothesis that relatively minor alterations in AGO-mediated regulation might be difficult to detect in a background of more robust distal regulation. Mutations were made as indicated in Figure 3.14 to attempt to disrupt the most likely AGO-miRNA seed sites as predicted by AGO-miRNA sequencing, with the expectation that derepression would only be evident in MDA-231 where endogenous AGO binding occurs. Mutations in the miR-28-5p seed site were used as a negative control since this miRNA is cloned at consistently low levels in all cell lines, representing less than 0.015% of the total AGO-miRNA population.

Pilot experiments using 50ng of vector showed no significant change in luciferase activity with any mutant constructs in either MDA-231 or 4175LM2 cells (data not shown). Given the relatively low expression of the miRNAs whose seeds are represented in the differential binding sites, the amount of transfected plasmid was decreased to 5ng to attempt a more stoichiometric relationship between target and endogenous miRNA. Transfections were carried out in sextuplet in 96-well plates using either full length (Figure 3.15A) or short HMGA1 3'UTR (Figure 3.15B) and are shown normalized to empty vector control.

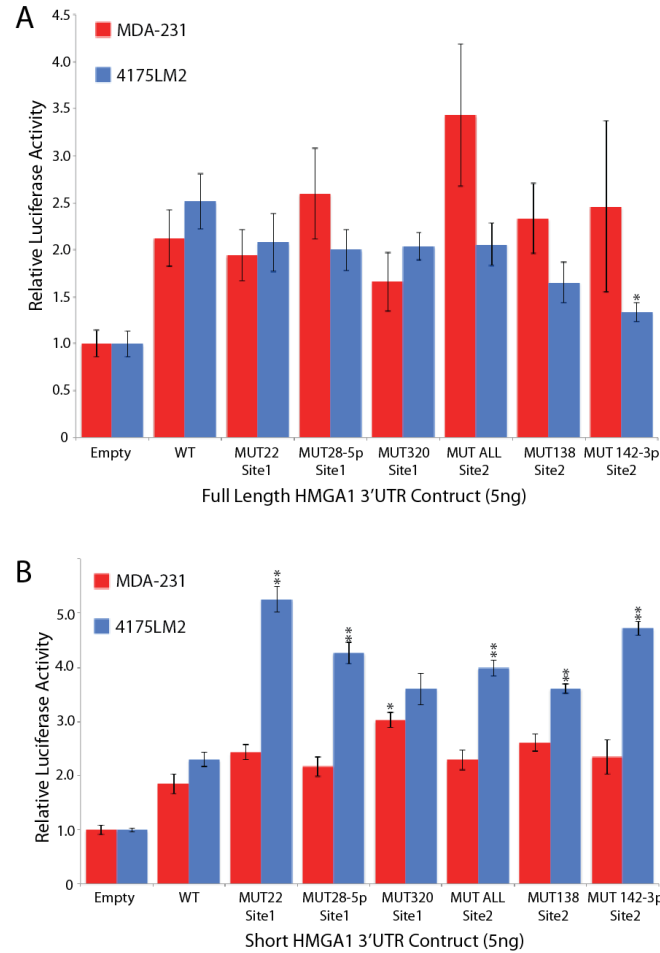


FIGURE 3.15: HMGA1 3'UTR Dual luciferase reporter assays in MDA and LM2 cells reveal unexpected differential miRNA regulation. Reporter constructs were transfected into MDA-231 and 4175LM2 non-reporter cell lines and luciferase activity assayed at 30hrs. Relative luciferase activity is shown normalized to empty vector, mutants are as outlined in Figure X. **A:** Relative expression of luciferase fused to full length HMGA1 3'UTR is unaffected by single seed mutations in site 1 and 2 in both MDA and LM2 cells. Combinatorial mutations of miR-142-3p and miR-320 seeds lead to increase in luciferase activity in MDA cells, although this increase failed to reach significance (MUT ALL Site2, $p=0.089$). Mutations in the seeds miR-138 or miR142-3p of site 2 (but not both) lead to repression in LM2 cells. **B:** 3' UTR reporter assays using a short fragment of HMGA1 spanning only the region of differential binding reveal significant de-repression in MDA cells transfected with miR-320a mutant, while unexpectedly all mutations lead to significant de-repression in LM2. $N=10$, p -values calculated by t -test compared to WT in that cell line, * $p \leq 0.05$ ** $p \leq 0.01$, error bars = s.e.m.

As anticipated, mutations in seed sequences in the differentially bound AGO-mRNA cluster region of the full length HMGA1 3'UTR had no effect on relative luciferase activity 4175LM2 cells when compared to WT. In MDA-231 cells, no significant derepression was seen in luciferase activity with mutant full length 3'UTRs, but a trend towards derepression in mutants with combined abrogated miR-142-3p and miR-138 seed sites was seen in multiple replicate experiments (MUT ALL site 2), suggesting that combinatorial loss of AGO regulation of these two miRNAs may influence HMGA1 expression levels to a small degree (Figure 3.15A).

In the context of the short HMGA1 3'UTR, mutations in the miR-320a/b/c seed sequence lead to significant increase in luciferase activity in MDA-231 cells. Surprisingly, in the context of the shorter HMGA1 3'UTR fragment significant derepression of all mutant constructs was observed in 4175LM2 cells, contrary to the predicted lack of AGO-regulation in these cells (Figure 3.15B). This would suggest that deletion of the distal region of the HMGA1 3'UTR allows for 4175LM2-specific derepression of the luciferase reporter construct independent of the seed mutated, and therefore, that loss of AGO binding on the HMGA1 3'UTR in metastatic cells may result from a decrease in accessibility of the transcript for AGO rather than being indicative of a loss of a specific miRNA.

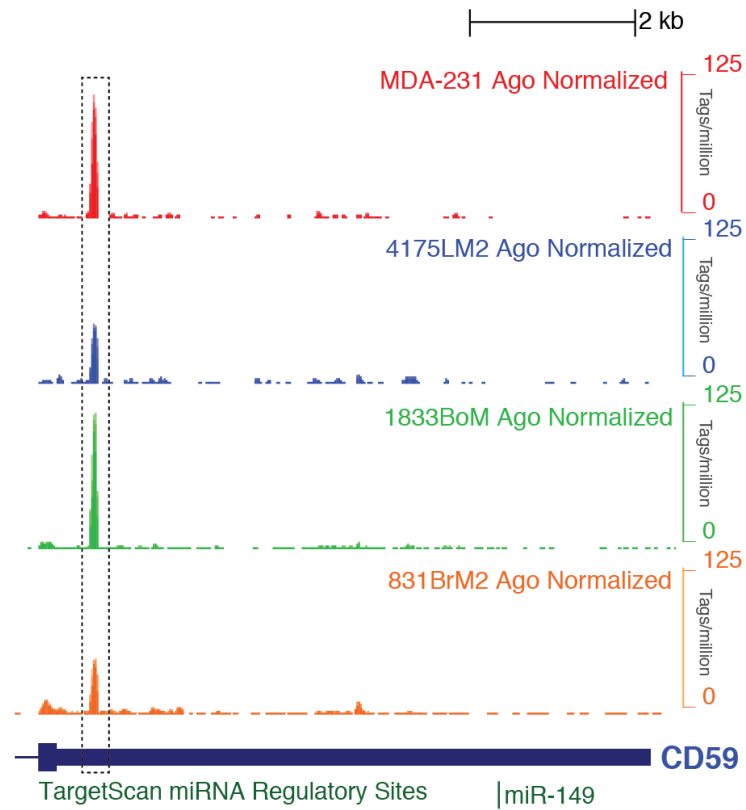


FIGURE 3.16: The 3'UTR of CD59 shows differential binding of AGO in single location with reduced binding in both 4175LM2 and 831BrM2 cells. Normalized AGO binding maps a single peak of AGO binding in the 3'UTR of CD59 that is decreased 1.6-fold in 4175LM2 ($p=0.002$) and 1.5-fold in 831BrM2 cells ($p=0.0003$), but is unchanged in 1833BoM2 cells (0.8-fold, $p=0.05$) relative to MDA-231 cells.

Differential AGO binding in the 3'UTR of CD59

In order to simplify the search for metastasis-associated loss of AGO binding, the criteria for selection was relaxed to identify clusters with significant enrichment in binding in any two metastatic sub lines compared to MDA-231, where AGO binding occurred predominantly in a single location in the 3'UTR. Complement regulatory protein 59 (CD59) was identified as having a single peak of AGO binding in the 3'UTR that decreased 1.6-fold in 4175LM2 ($p=0.002$) and 1.5-fold in 831BrM2 cells ($p=0.0003$) relative to MDA-231, but is relatively unchanged in 1833BoM2 cells (0.8-fold, $p=0.05$) (Figure 3.16). Previous microarray data and subsequent Q-PCR validation (Figure 3.17A) confirmed that CD59 transcript levels are approximately two-fold higher in 4175LM2 than MDA-231, consistent with a loss of negative regulation. FACS analysis of CD59, a cell surface protein, confirmed that this increase in mRNA translated into increased CD59 protein expression in 4175LM2 but not 831BrM2 relative to MDA-231 (Figure 3.17B).

The ~80bp footprint of AGO binding on the 3'UTR of CD59 contains seed sequences for multiple miRNAs, of which miR-125a-5p/a and miR-183 are the most highly represented in the combined AGO-miRNA HITS-CLIP profile (Figure 3.18A, ranked 13th and 39th in MDA-231 respectively). Neither miR-125a-5p/b nor miR-183 show significant changes in AGO-miRNA tag counts between cells lines. The regulation of CD59 3'UTR by miR-183 was assessed in a 326nt CD59 3'UTR fragment spanning the AGO binding site (chr11:33,687,643-33,687,968).

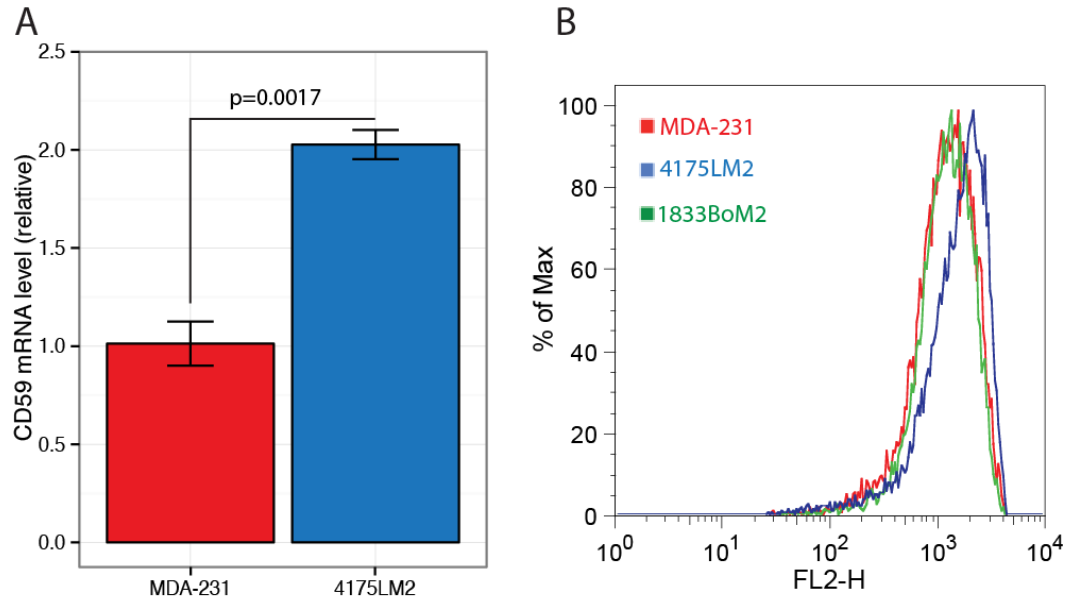


FIGURE 3.17: Decreased AGO 3'UTR binding in CD59 correlates with increase in both mRNA and protein levels in 4175LM2 cells. **A:** Q-PCR for CD59 reveals a two-fold increase in mRNA levels in LM2 cells consistent with previous microarray studies. (n=3, normalized to TBP, t-test, error bars = s.e.m). **B:** FACS staining for CD59 reveals increased protein levels in 4175LM2 but not 1833BoM2 cells.

Co-transfection of this WT sequence construct into 4175LM2 cells with 50nM of miR-183 mimic or a negative control mimic had no effect on relative luciferase activity (Figure 3.18B) suggesting that miR-183 does not mediate the AGO-mRNA interactions identified by AGO HITS-CLIP. Further mimic transfections need to be carried out in order to establish the identity of the miRNA targeting AGO to CD59 at this site. Alternatively, a modification of the HITS-CLIP protocol that utilizes intermolecular ligation to produce chimeric AGO-miRNA-mRNA, termed CLASH (Helwak et al., 2013), may be useful to identify with high confidence the miRNA responsible for targeting AGO to this site.

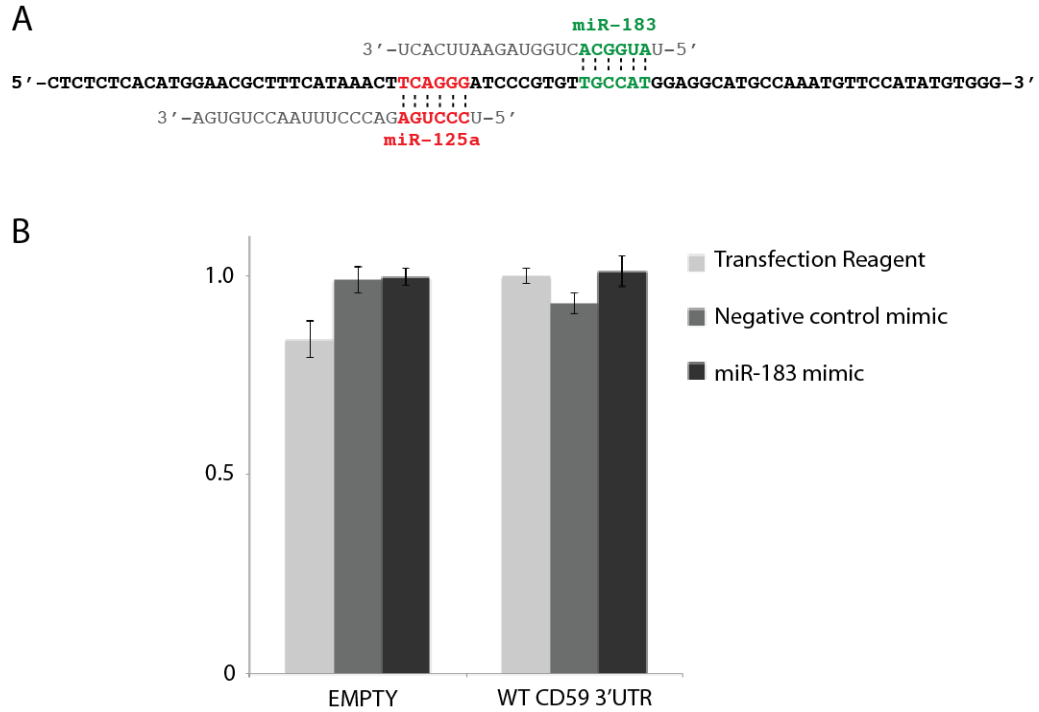


FIGURE 3.18: The differentially bound site in CD59 3'UTR contains seed sequences for miR-183 and miR-125a-5p/b, but it not regulated by miR-183. **A:** The AGO footprint region of the CD59 3'UTR with seed matches to miR-183 and miR-125a-5p highlighted. **B:** Dual luciferase reporter assays show that miR-183 does not regulate the full-length WT CD59 3'UTR. Transfections were done in triplicate with 50nM of miR-183 mimic or a negative control mimic that does not target any known human transcripts, or with plasmid and transfection reagent alone. Error bars = s.e.m.

DISCUSSION

Altered miRNA expression is a common hallmark of cancer, with general down regulation miRNAs compared to normal tissues occurring in patterns that can be used as classifiers of tumor type (Lu et al., 2005). Conversely, gain of miRNA expression can be oncogenic or tumor suppressive depending upon the specific mRNA targets co-expressed and the cellular context (reviewed in Lujambio and Lowe, 2012). An increasing body of work has implicated altered miRNA regulation both in breast cancer metastasis suppression and promotion (Wang and Wang, 2012), with miR-335 and miR-126 known to regulate metastasis in the model system interrogated here (Png et al., 2011, Tavazoie et al., 2008). By applying the AGO HITS-CLIP methodology in this well-established and molecularly characterized model of breast cancer metastasis with known miRNA regulatory components, we aimed to identify both miRNAs differentially bound to AGO with respect to metastatic potential and the resulting sites of altered AGO-mRNA regulation on a genome-wide scale.

This study has generated a comprehensive binding map of endogenous human AGO bound to endogenously expressed miRNA and cognate mRNA targets in the human metastatic breast cancer cell line, MDA-MB-231, and its aggressively metastatic sub-lines with tissue tropism to lung, bone and brain (Bos et al., 2009, Kang et al., 2003, Minn et al., 2005a). This is in contrast to previously published human AGO-mRNA data sets that have been generated after transfection with exogenous tagged-Argonaute or miRNA mimics (Chi et al., 2009, Hafner et al., 2010,

Helwak et al., 2013), or were done in virally infected cells (Riley et al., 2012). The transcriptome-wide AGO-mRNA binding sites described here provide a valuable reference set for researchers; providing independent and direct biochemical evidence of robust, endogenous AGO-mRNA interactions to supplement experimental validation of miRNA mediated effects on predicted target mRNA and (reporter) protein levels. As an example, several reports have implicated miR-29a in the regulation of SPARC in both breast cancer cells (MCF-7 (Wang et al., 2012), MDA-MB-435 (Gerson et al., 2012)) and osteoblasts (Kapinas et al., 2009). The greatest relative AGO occupancy on the SPARC 3'UTR (Figure 3.10) occurs in a cluster comprised of two peaks directly corresponding to two TargetScan predicted, conserved miR-29 seed sequences. Given that miR-29a is one of the most abundantly CLIPed miRNA in all four cells lines, and the consistent finding that the most frequently cloned AGO-miRNAs have significantly more seed pairs within the AGO-mRNA binding sites than the least cloned miRNAs (Chi et al., 2009, Zisoulis et al., 2010), binding at this site is highly predictive of miR-29a regulation of SPARC in lung and brain metastatic sub lines on MDA-231.

The most frequently CLIPed miRNA in all four cell lines, miR-27a, has previously been identified as a potential oncomiR in gastric cancer (Liu et al., 2009) and MDA-MB-231 cells (Mertens-Talcott et al., 2007), and high levels are associated with poor prognosis in invasive breast cancers (Tang et al., 2012b). The abundance of the top 20 AGO-miRNA derived from HITS-CLIP in MDA-231 cells was found not to correlate with

previously published high-throughput sequencing profiles of MDA-MB-231 total miRNA (data not shown, (Mayr and Bartel, 2009)). This is in contrast to previous reports of high degree of correlation between Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) derived AGO-miRNA counts and the total miRNA abundance profile in HEK293 cells (Hafner et al., 2010). PAR-CLIP is a HITS-CLIP variation that utilizes the incorporation of 4-thiouridine (4SU) into transcripts of cultured cells, which upon UV 365nm irradiation causes T to C transitions in the cloned CLIP library. The discrepancy between published miRNA profiles and the AGO-miRNA HITS-CLIP profile in MDA-231 may reflect true biological differences between absolute expression levels and preferential incorporation into RISC, or could result from technical differences in cloning (Jayaprakash et al., 2011), cell stocks or culture conditions; for example, increasing cell density is known to lead to widespread increase in miRNA abundance (Hwang et al., 2009). Parallel high-throughput miRNA sequencing of the cell lines used in this study will be required to distinguish between these two possibilities.

Assessment of the AGO-miRNA profiles obtained from MDA-231, 4175LM2, 1833BoM2 and 831BrM2 cells revealed a striking degree of correlation within the 30 most frequently CLIPed miRNAs, with very few having statistically significant differences in the HITS-CLIP profiles. This was unanticipated given the number of miRNA previously identified as differentially expressed in these cells with respect to both metastatic

ability and tissue tropism (Tavazoie et al., 2008). The virtual absence from the MDA-231 AGO-miRNA profile of the previously validated anti-metastatic miR-335 (0.0009 % of miRNA) and miR-126 (not cloned) indicates that the previously reported correlation between miRNA expression and AGO-miRNA CLIP profiles may not extend to minority miRNA species. This represents an unforeseen limitation on the use of differential AGO HITS-CLIP to determine on a genome-wide scale the mRNA targets of known metastasis-inhibitory endogenous miRNAs in MDA-MB-231 cells.

Comparative analysis of the HITS-CLIP derived AGO-miRNA profiles identified one miRNA, miR-142-3p, with read counts significantly decreased in 4175LM2 and increased in both 1833BoM2 and 831BrM2 relative to MDA-231. Expression of miR-142 has previously been reported to be highly specific to hematopoietic cells (Landgraf et al., 2007), with high levels of expression in naïve CD8⁺ T cells (Wu et al., 2007) and a reported role in the regulation of monocyte terminal differentiation (Lagrange et al., 2013). Recent studies have implicated miR-142-3p in tumor suppression of both colon cancer and hepatocellular carcinoma, while mutations in both strands of pre-miR-142 have been found in diffuse large B cell lymphoma (Kwanhian et al., 2012, Shen et al., 2013, Wu et al., 2011a). While miR-142-3p is a relatively minor miRNA in all four cell lines, small changes in miRNAs expressed at low levels have been shown to have profound functional consequences, for example the anti-metastatic miR-335, miR-126 and miR-206 (Tavazoie et al., 2008) and the

miR-183/96/182 cluster in light adaption in the retina (Krol et al., 2010). The differential AGO-CLIP profiles obtained for miR-142-3p identify it as an interesting candidate for further investigation to determine whether it functions as a cell type specific metastasis suppressor (4175LM2) or activator (831BrM2 and 1833BoM2), and whether differential expression clinically correlates with the site of distal metastases in breast cancer patients.

The AGO-miRNA profiles of three additional miRNAs represented in top 30 most abundant seed families were found to be specific decreased in 4175LM2 cells; let-7c, miR-99a and miR-20b. miR-99a and let-7c, along with miR-125b-2, are expressed as a cluster derived from an intron of the non-coding RNA, C21orf34. The miR-99a/let-7c/125b-2 cluster is located on chromosome 21 in a region commonly deleted in both primary breast (Ohgaki et al., 1998) and lung cancers (Calin et al., 2004, Yamada et al., 2008). Interestingly, miR-125b is transcribed from two locations within the genome and while the mature miR-125b AGO-miRNA counts were unchanged between cell lines, miR-125b-2* was significantly lower in 4175LM2 than all other cell lines (LM2/MDA \log_2 FC=-2.58 $p=0.024$, BoM/LM2 \log_2 FC=4.0 $p=1.61e^{-4}$, BrM/LM2 \log_2 FC=3.5 $p=6.2e^{-5}$). This strongly suggests specific decrease in the incorporation of all three members of the miR-99a/let-7c/125b-2 cluster into AGO in lung metastatic cell lines, identifying this cluster as a potential novel regulator of lung specific metastasis.

Given the unexpectedly similar AGO-miRNA profiles of MDA-231, 4175LM2, 1833BoM2 and 831BrM2 and the relatively low expression of those miRNA that showed significantly altered AGO-miRNA profiles, the potential scope for seed-directed searches for alternatively regulated mRNAs were limited. This was further compounded by the potential for redundant targeting by other seed family members, as the seed sequences of individual miRNA showing significant differences in AGO-miRNA profiles were still abundantly represented in all cell lines. For example, let-7c has been demonstrated to negatively regulate the oncogene HMGA2 in uterine leiomyoma via multiple target sites in the 3'UTR (Klemke et al., 2010). However, this regulation has also been demonstrated when using exogenous let-7a (Mayr et al., 2007), mixtures of let-7b/d/g (Guo et al., 2013) or let-7b/e (Lee and Dutta, 2007), and by family member miR-98 in head and neck squamous cell cancer (Hebert et al., 2007). Accordingly, the multiple robust AGO-mRNA peaks corresponding to HMGA2 3'UTR let-7 seed sites are unchanged in 4175LM2 when compared to the other four cell lines (data not shown).

Therefore, differential AGO-mRNA binding was assessed without prior assumption as to the miRNA involved in targeting, with the simplistic view that commonly gained or lost clusters in aggressively metastatic sub-lines would identify deregulated targets related to breast cancer biology. This approach allowed for inclusion of both miRNA-driven binding changes resulting from miRNA expression level changes, and non-miRNA driven effects such as shortening of 3'UTRs via

alternative splicing or polyadenylation site usage. Unexpectedly, very few AGO-mRNA binding sites were either commonly gained or lost in the highly metastatic sub-lines of MDA-231, however those transcripts identified in this study as having altered AGO-mRNA binding sites with corresponding changes in transcript level include several with strong links to metastasis.

Gain of AGO-mRNA binding was identified in the 3'UTR of TUSC2, a gene that is located in a region of the genome frequently deleted in breast and lung cancer (3p21.3 (Zabarovsky et al., 2002)). TUSC2 is pro-apoptotic when ectopically expressed in non-small cell lung cancer cells harboring 3p21.3 deletions (Ji et al., 2002, Kondo et al., 2001), and ectopic re-expression via nanoparticle delivery is currently in phase I clinical trials for the treatment of recurrent or metastatic lung cancer (Lu et al., 2012). Previously described miRNA regulation of TUSC2 by miR-93 and miR-98 corresponds with the AGO-mRNA binding peaks commonly found in all four breast cancer cell lines (Du et al., 2009). Interestingly, deletion of the distal portion of the TUSC2 3'UTR, the region containing metastatic sub-line specific AGO-mRNA *de novo* binding, leads to increased TUSC2 protein level in human lung cancer cells (Lin et al., 2011) suggesting that these sequences contain potent negative regulators of translation.

CD59 is a membrane-bound protein that protects against complement-mediated lysis by interacting with C8 and C9 to preventing membrane attack complex (MAC) formation (Meri et al., 1990). Up

regulation of CD59 has been described in multiple malignant cell types (Fishelson et al., 2003) including in breast cancer, where CD59 has a role in both resistance to complement-mediated lysis and HBXIP-mediated proliferation and migration (Cui et al., 2012). CD59 is expressed throughout the human respiratory tract (Varsano et al., 1995), with increased levels reported in lung cancer cell lines (Varsano et al., 1998). Decreased AGO binding, the correlated increase in both mRNA and protein levels of CD59 seen in 4175LM2 cells may be beneficial in providing protection against complement attack during colonization and out growth in the lung metastatic niche. It will be interesting to determine whether knockdown of CD59 in these cells negatively influences their ability to form metastatic lesions in the lung.

Both HMGA1 and HMGA2 are oncofetal proteins, with high levels of expression found during embryogenesis and several tumor types and undetectable or very low expression in differentiated adult tissues (Chiappetta et al., 1996, Zhou et al., 1995). While HMGA proteins can function as anti-proliferative components of cellular senescence, they can also promote proliferation depending on cellular context (Narita et al., 2006) and are overexpressed in multiple human malignancies (Fedele and Fusco, 2010). Overexpression of HMGA1 in MCF-7 cells leads to enhanced tumor formation and metastasis (Liu et al., 1999, Reeves et al., 2001) with reciprocal decrease in both orthotopic tumor growth and lung metastasis found with HMGA1 knockdown in MDA-231 cells (Shah et al., 2013).

While the posttranscriptional regulation of HMGA2 by let-7 is well documented and corroborated by the robust AGO HITS-CLIP binding data generated in this study, the regulation of HMGA1 by miRNAs in breast cancer is less clear. AGO-mRNA binding is seen in regions corresponding to previously validated miR-26, miR-15/16 and let-7 sites (Figure 16, (Palmieri et al., 2012)), however none of the previously validated regulatory miRNA seed sites found in the literature map to the region of differential AGO binding identified by HITS-CLIP. Mutation of the most likely candidate miRNA seeds within these sites (based on AGO-miRNA profiling) failed to lead to consistent significant derepression in MDA-231. It remains to be tested whether other relatively minor miRNA are able to regulate HGMA1 at the sites identified by HITS-CLIP. However, the seemingly anomalous result whereby removal of the distal portion of the 3'UTR allows for mutation-mediated derepression of HGMA1 reporter constructs specifically in 4175LM2 would suggest that other *cis*-regulatory elements may be responsible for the loss of AGO-binding seen in these cells under endogenous full length 3'UTR conditions. Such antagonistic regulation has been described in zebrafish, where *cis*-regulatory elements in the 3'UTR of *nanos1* are sufficient to inhibit ubiquitously expressed miR-430-mediated deadenylation and translational repression in primordial germ cells but not the soma (Mishima et al., 2006).

Identification of the miRNAs involved in targeting AGO to the sites of differential binding described in this chapter has proven challenging.

Potential regulatory miRNAs were selected for mutational reporter assay analysis based on both the seed sequences within the AGO footprint of differential binding and the abundance of the complementary miRNA in the AGO-miRNA profile. While miRNAs known to mediate metastasis suppression were below the threshold of AGO-miRNA HITS-CLIP detection, it is not clear how this lack of detectable miRNA binding translates in the AGO-mRNA profile. Rare miRNA targeting may result in robust AGO-mRNA clusters under certain conditions, for example, if regulated targets were present at high concentration and/or the miRNA in question regulated relatively few transcripts thereby increasing the relative concentration of a rare AGO-miRNA at the target sites. Such binding may help to explain the “orphan” clusters found in previous AGO HITS-CLIP experiments that do not contain seed matches to the top 30 miRNAs in the AGO-miRNA profile (Chi et al., 2009).

The AGO-mRNA footprints derived from HITS-CLIP are valuable for delimiting the sequence space in which to search for miRNA seed matches, however increasing HITS-CLIP-derived evidence points to substantial non-canonical yet functional AGO-miRNA-target pairing. For example, it has been recognized for some time in *C.elegans* that the *lin-4::lin-14* interaction accommodates a C-bulge (Ha et al., 1996, Wightman et al., 1993), and that the *let-7::lin-41* interaction involves a G:U wobble (Vella et al., 2004), with both of these direct interactions confirmed by *alg-1* HITS-CLIP (Zisoulis et al., 2010). Analysis of multiple AGO HITS-CLIP libraries indicates a high degree of functional miRNA-mediated AGO-mRNA

regulation at non-canonical sites including various seed-like motifs (G-bulges at position 5-6 (Chi et al., 2012), single mismatches (Loeb et al., 2012)) and interactions outside of the seed that are complementary to the middle and 3' end of the miRNA (Helwak et al., 2013). Furthermore, it has been noted that AGO binds mRNA preferentially in G-rich motifs in a similar genomic distribution in mouse embryonic stem cells with or without mature miRNAs (Leung et al., 2011). These alternate rules of regulation add a further layer of complexity to the search for miRNAs responsible for targeting AGO to the *de novo* clusters identified in MDA-231 and its derivatives. The newly developed protocol CLASH, in which intermolecular ligation is used to drive the formation of chimeric miRNA-target mRNA AGO HITS-CLIP tags (Helwak et al., 2013), may be useful for identifying the miRNAs involved in regulating the differentially AGO bound targets identified in this study.

Close inspection of the previously published expression data comparing MDA-231 and 4175LM2 (Minn et al., 2005a) or 831BrM2 (Bos et al., 2009) revealed an additional potential variable regarding regulation of AGO-mRNA binding, namely altered expression of other RNA binding proteins. Muscleblind 2 (MBNL2) is among the downregulated genes identified in 4175LM2 cells (Minn et al., 2005a), with CUGBP2/CELF2 and ELAVL4/HuD being upregulated in 831BrM2 (Bos et al., 2009, Li et al., 2011a). Additionally, two-dimensional difference gel electrophoresis has identified hnRNPA2B1 as upregulated in independently derived brain metastatic clones of MDA-231 (Yoneda et al., 2001, Li et al., 2011b). Both

MBNL2 and CUGBP2 have been implicated in accelerated mRNA decay mediated by 3'UTR binding of targets (Du et al., 2010, Masuda et al., 2012), with recent HITS-CLIP data confirming that the majority of MBNL2 binding occurs in 3'UTRs (Charizanis et al., 2012). Conversely, hnRNPA2BP1 has been shown to stabilize long-lived mRNA transcripts in MDA-231 cells via interactions with RNA structural motifs in their 3'UTRs (Goodarzi et al., 2012).

While no direct link between the 3'UTR binding of these proteins and AGO-mediated regulation has been investigated to date, the potential for ELAV-modulation of miRNA target site accessibility is known. Jacobsen et al. (2010) have shown that the binding motif for ELAVL4 is the most enriched motif in mRNAs with altered expression levels following small RNA transfection in HeLa cells. ARE motifs were found to be significantly enriched in regions ~50nt upstream of endogenous miRNA targets previously defined by AGO HITS-CLIP, correlating with both attenuated and augmented transfected miRNA-dependent destabilization. Significantly, overlapping genome-wide patterns of HuR and AGO binding have been described using PAR-CLIP in HEK293 cells (Mukherjee et al., 2011) and with endogenous proteins via HITS-CLIP in resting and activated human T-cells (Darnell Lab, unpublished data), suggesting a potential role for agonistic or antagonistic HuR regulation of AGO binding on common target transcripts. Indeed examples of both types of interaction exist. HuR binding is required for let-7 mediated c-Myc repression (Kim et al., 2009a) and has been shown to relieve miR-122

repression of CAT-1 mRNA (Bhattacharyya et al., 2006). Comparing AGO-mRNA binding across cell lines where potential “unseen” RBP modulators of transcript stability and AGO targeting are known to be differentially expressed further complicates the seed-directed approach towards validation of differential AGO-mRNA binding.

Perspectives

By pairing differential AGO-mRNA HITS-CLIP analysis with expression data, we have identified potential for miRNA-mediated regulation of metastatic organotropism, with miR-142-3p and the miR-99a/let-7c/miR-125-b2 cluster specifically downregulated in lung metastatic cell lines. We have shown that it is possible to identify transcripts in which gain or loss of AGO binding correlates with decrease or increase in steady state mRNA levels and, in the case of CD59, protein levels. Those transcripts identified as having *de novo* AGO-mRNA binding have well documented roles in cancer biology and metastasis. While the particular miRNA involved at the identified differential AGO binding sites are currently unknown, it will be interesting to investigate the function of metastasis-specific *de novo* AGO-mRNA binding in the promotion or inhibition of metastasis via transfection with antisense oligonucleotides that block AGO binding to specific target mRNA. For simplicity we have focused on identifying differential binding common to the aggressively metastatic phenotype in 3'UTRs, however the genome-wide nature of the data presented allows for future analysis of binding in

other regions of transcripts and in identifying differentially regulated targets with respect to tissue tropism.

CHAPTER 4: TRANSCRIPTOME-WIDE TARGET IDENTIFICATION OF THE ANTI-METASTATIC miR-335

INTRODUCTION

Deregulation of miRNA expression in tumors can be accomplished by multiple mechanisms. Many miRNA loci are located in fragile sites, common breakpoint regions, and cancer-associated sites of loss of heterozygosity or amplification (Calin et al., 2004, Zhang et al., 2006). miRNAs may be epigenetically silenced by promoter DNA methylation (Choudhry and Catto, 2011) and histone modifications (Scott et al., 2006). Furthermore, mature miRNA abundance can be affected by post-transcriptional regulation of miRNA processing and stability. For example, DNA-damaged induced up regulation of a subset of mature miRNAs is dependent upon p53 interaction with Drosha (Suzuki et al., 2009), while estrogen receptor-alpha can suppress the maturation of miR-125a and miR-145 in breast cancer cells (Yamagata et al., 2009).

Using the xenograft model outlined in Chapter 3, (Tavazoie et al., 2008) have identified three endogenously expressed human miRNAs, miR-126, miR-206 and miR-335, that robustly suppress breast cancer metastasis to lung and bone, and whose expression levels negatively correlate with patient metastasis free survival. Loss of miR-206 expression has been independently associated with lymph node metastasis and poor prognosis in a cohort of breast cancer patients (Li et al., 2013), and has been reported to enhance metastatic spread of rhabdomyosarcoma

(Missiaglia et al., 2010), endometrioid adenocarcinoma (Chen et al., 2012), and lung cancer cells (Wang et al., 2011c). Loss of miR-126 in breast cancer leads to increased proliferation and number of metastatic nodules in lung, indicating a role in the regulation of metastatic initiation and colonization. Interestingly, de-repression of miR-126 targets IGFBP2, MERTK and PITPNC1 leads to enhanced endothelial cell recruitment and angiogenesis (Png et al., 2012) indicating a non-cell autonomous role for miR-126 in metastasis suppression. Loss of both miR-126 and miR-126* expression has been shown to enhance metastasis in a murine model of breast cancer via up regulation of the endothelial cell recruiting chemokine SDF-1 α (Zhang et al., 2013).

The miR-335 locus is located in the second intron of mesoderm specific transcript (MEST). Decrease in miR-335 expression is associated with enhanced invasiveness of breast cancer (Heyn et al., 2011, Png et al., 2011, Tavazoie et al., 2008), gastric cancer (Xu et al., 2012) and neuroblastoma (Lynch et al., 2012). For miR-335, both genetic loss and promoter hypermethylation of the MEST host gene has been described in clinical breast cancer samples (Png et al., 2011) and hepatocellular carcinoma (Dohi et al., 2013), with loss of expression correlating with higher rates of distal metastasis formation. In contrast, miR-335 expression is pro-oncogenic in malignant astrocytoma where it promotes growth and invasion (Shu et al., 2011). Expression of miR-335 has been proposed to control proliferation in U2OS cells in a p53-dependent manner via direct retinoblastoma 1 (Rb1) targeting (Scarola et al., 2010), while this same Rb1

interaction is proposed to enhance meningioma cell growth (Shi et al., 2012). MDA-231 cells express a non-functional mutant p53 (Petitjean et al., 2007), and therefore this pathway is unlikely to influence metastatic potential in these cells or its derivatives. An additional role for miR-335 in the regulation of stem cell self-renewal has been proposed. Low miR-335 expression in breast cancer stem cells promotes cell growth (Polytarchou et al., 2012), while up-regulation of miR-335 in mouse epithelial stem cells promotes differentiation by targeting Oct-4 and retinoblastoma protein (Schoeftner et al., 2012). Conversely, miR-335 has been found at high levels in human mesenchymal stem cells where it both prevents differentiation and inhibits motility (Tome et al., 2011).

miR-335, miR-126 and miR-206 were identified as potent endogenous inhibitors of metastasis by comparative microarray profiling of the relatively weakly metastatic MDA-231 cell line and its aggressively metastatic sub-lines with tropism to lung and bone (Tavazoie et al., 2008). While the fold changes in expression level of these miRNAs were large, between 50-70% for miR-335, the absolute levels of each miRNA were low in the parental cells. Consequently, these miRNAs were not amenable to the differential AGO HITS-CLIP analysis carried out on MDA-231 and 4175LM2 cells where miR-335 and miR-335* together comprise <0.01% of all MDA-231 crosslinked AGO-miRNA clones. Therefore to globally identify miR-335 targets, we made use of previously characterized lung metastatic cell lines with restored miR-335 expression and reduced metastatic ability. These cells were created by cloning an approximately

500bp region of the second intron of MEST (spanning the miR-355 genomic location) into a pMSCV-Blasticidin vector, allowing stable integration and overexpression of pre-miR-335 under the control of a CMV promoter (miR-Vec, (Tavazoie et al., 2008, Voorhoeve et al., 2007). This system requires the normal cellular processing machinery to produce both pre- and mature- miRNA guide strand so maintaining any endogenous regulation of processing and miRNA stability.

Given the multiple roles described for miR-335 in metastasis progression and the strong negative correlation between expression levels and clinical metastasis-free survival, miR-335 was selected for further analysis using AGO HITS-CLIP. By mapping AGO-mRNA interactions in the presence and absence of miR-335 expression, we aimed to generate a comprehensive catalog of directly regulated miR-335 target mRNAs to more fully understand its anti-metastatic action.

RESULTS

AGO HITS-CLIP in 4175LM2 Control and 4175LM2 pre-miR335 overexpressing cells

Argonaute HITS-CLIP was carried out in triplicate as described in Chapter 3 using stable 4175 LM2 cell lines expressing either a control, non-targeting hairpin designed against inhibitor of DNA binding 1, ID1 (Minn et al., 2005a) or pre-miR-335 (Tavazoie et al., 2008), termed LM2 Control and LM2 335OE, respectively. Overexpression of miR-335 and miR-335* were confirmed by Q-PCR before HITS-CLIP experiments were undertaken, with both detected at ~120 fold higher levels in LM2 335OE compared to LM2 Control (Figure 4.1). The resulting high-throughput sequencing data was processed and filtered as in Figure 3.5, with the number of miRNA and unique mRNA tags obtained from each replicate library outlined in Table 4.1. Read depth between the two experiments was more equally distributed than previously described for MDA-231 cells and metastatic sub lines, with ~1.7 million unique mRNA tags obtained for LM2 Control cells and ~1.4 million unique mRNA tags from LM2 335OE.

Characterization of the AGO-miRNA profiles of LM2 control and LM2 335OE

AGO-miRNA data sets were analyzed to determine whether overexpression of pri-miR-335 disrupts the global miRISC profile of LM2 cells. Read counts per miRNA tallied as previously described, miRNAs grouped by 7-mer seed family and the percentage of reads calculated based on read depth in that experiment. The average percentage per seed

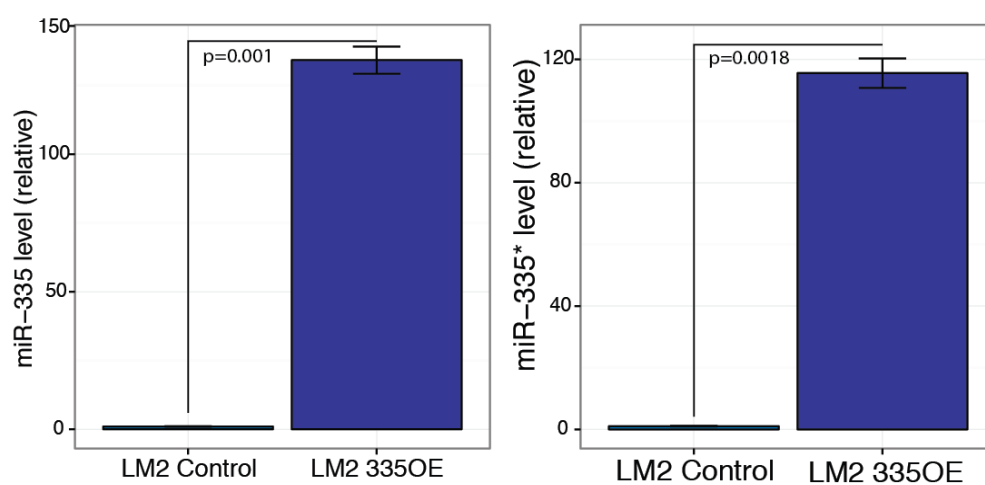


FIGURE 4.1: Verification of miR-335/* over expression in stable cells lines. Q-PCR verification of miR-335 and miR-335* over-expression in LM2 335OE prior to AGO HITS-CLIP analysis. N=3, expression levels normalized to Snord44, error bars= s.e.m, p-value calculated by t-test.

	LM2 CONTROL			LM2 335OE		
	A	B	C	A	B	C
Filtered & Adapter clipped	6,320,704	6,687,553	7,091,986	7,339,166	5,344,276	5,752,636
Mapped to hg18	3,462,271	2,903,604	3,025,865	4,355,941	3,099,865	3,300,299
miRNA removed	2,640,298	2,764,777	2,860,670	2,682,455	2,590,358	2,860,399
miRNA Tags	1,249,761	213,254	261,895	664,181	175,236	155,401
Unique mRNA Tags	541,992	472,405	663,330	405,315	469,006	546,861
	1,677,727			1,421,182		

Table 4.1: AGO HITS-CLIP sequencing results from triplicate LM2 Control and LM2 335OE cell lines. Read counts are shown for each data processing step as outlined earlier in Figure 3.5.

family was plotted in pair-wise scatter plots comparing LM2 Control and LM2 335OE with each other, and to the previously described 4175LM2 AGO-miRNA data set. Extremely high Pearson correlation coefficients were obtained when comparing all three data sets, indicating that neither the selective overexpression of pri-miR-335 nor the non-targeting control shRNA lead to a significant shift in the overall composition of miRNAs bound to AGO in 4175LM2 cells (Figure 4.2). Correlation coefficients were unaffected by removal of the top three most CLIPed miRNAs (data not shown).

*AGO HITS-CLIP reveals strand bias towards miR-335**

Plotting the percentage of miRNA CLIP tags from each LM2 cell line on a log scale revealed a clear increase in the representation of both miR-335 and miR-335* in LM2 335OE compared to both LM2 Control and 4175LM2 (Figure 4.3). The miRNA profiles of LM2 335OE and LM2 Control cells were compared using EdgeR software as described previously, with miRNAs having counts per million >2 in >4/6 HITS-CLIP AGO-miRNA profiles retained for analysis (Table 4.2). Surprisingly, while miR-335 showed a significant increase in the LM2 335OE AGO-miRNA profile, this fold change did not pass the FDR threshold ($\log_2FC=4.01$, $p=0.0025$, $FDR=0.12$). However miR-335*, the presumed pre-miR-335 passenger strand, had the greatest and most significant positive fold change of all AGO-miRNA ($\log_2FC=6.65$, $p=7.61E^{-15}$, $FDR=5.29E^{-12}$). Six miRNA showed significant decrease in the LM2 335OE miRNA profile

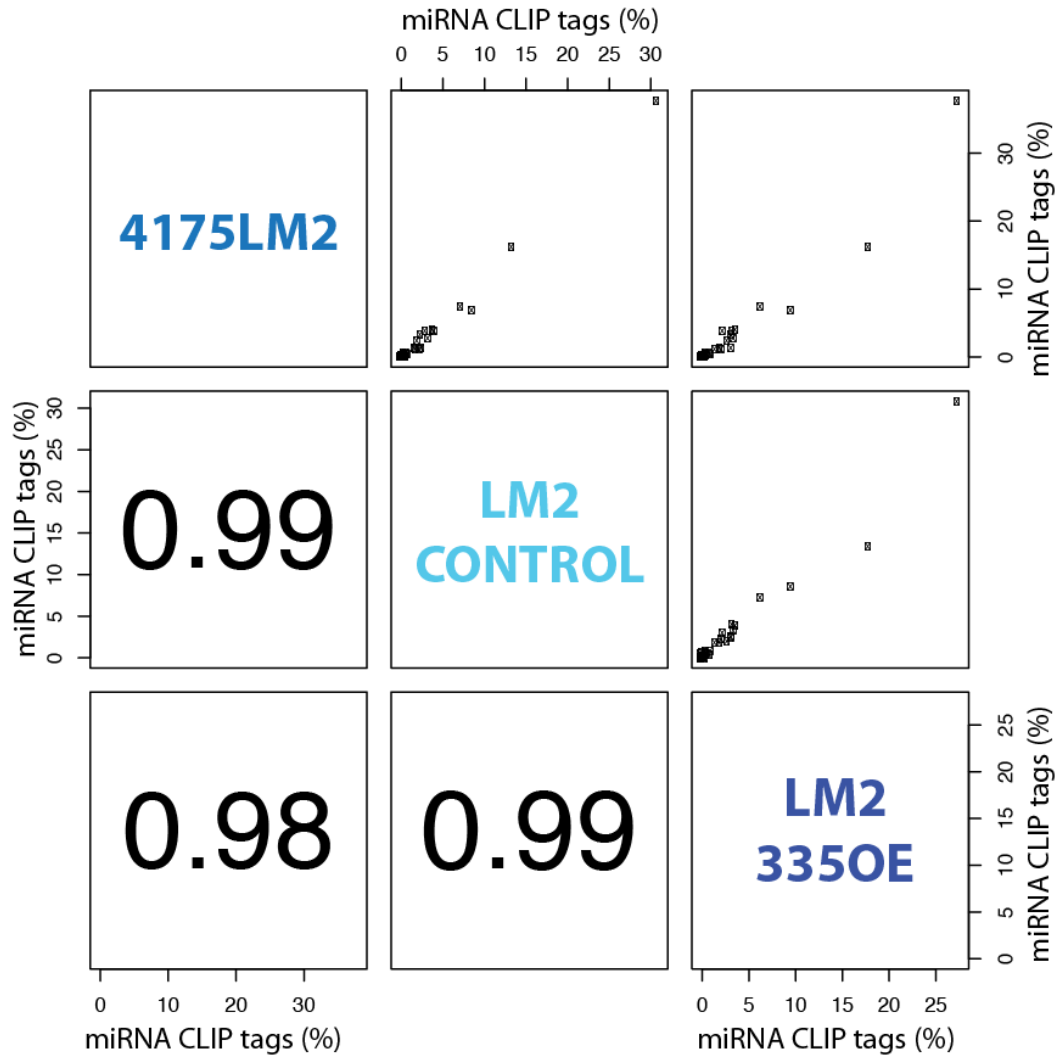


FIGURE 4.2: AGO-miRNA profiles in LM2 cell lines are unaffected by pri-miR-335 over expression. Expressed miRNAs ($BC \geq 2$ tags ≥ 10) common to 4175LM2, LM2 Control and LM2 335OE cells were compressed into seed families based on position 2-8 (7-mer) seed match. The percentage of reads per miRNA was calculated per library using the total number of reads mapping to miRNAs in that library as a denominator. Percentages were then averaged across replicates per cell line. Upper half: Pairwise scatter plots of miRNA family % in 4175LM2, LM2 Control and LM2 335OE. Lower half: absolute value of the Pearson correlation coefficient.

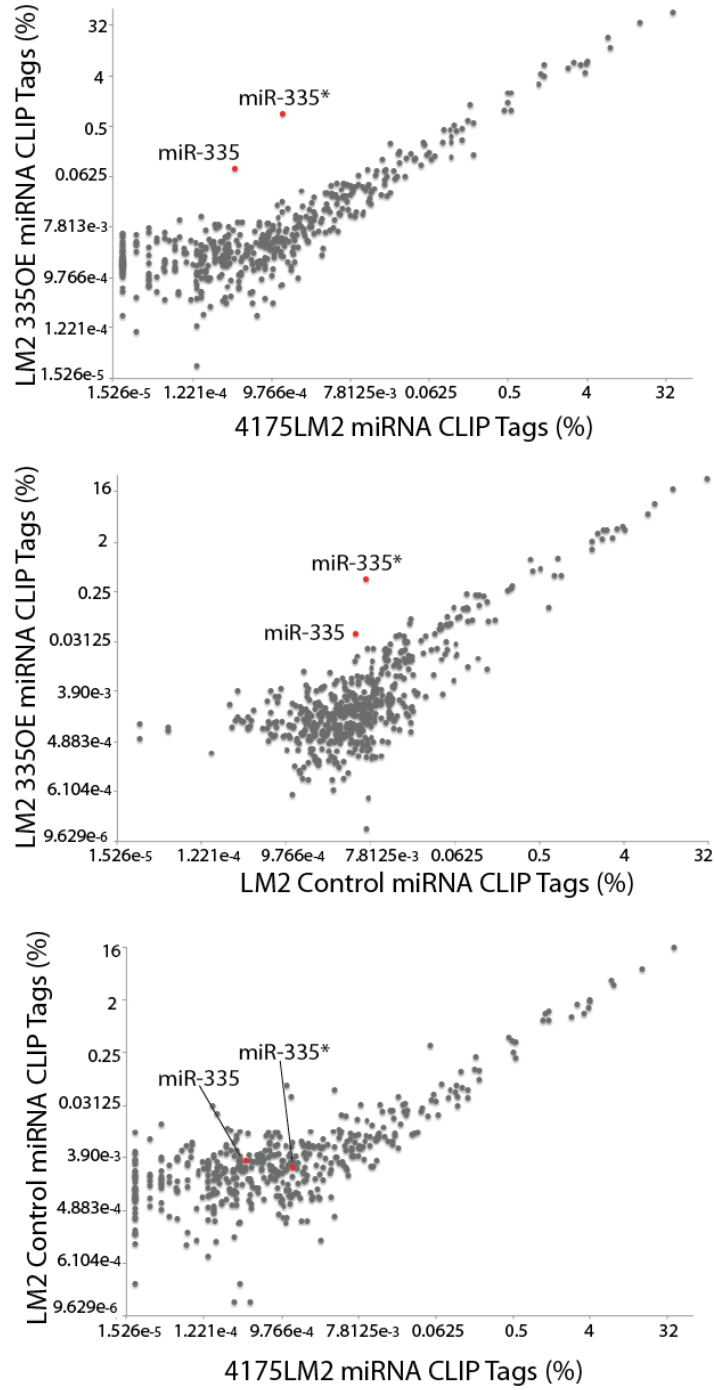


FIGURE 4.3: Over expression of pri-miR-335 leads to an increase in the percentage of AGO-miR-335 and AGO-miR-335* in LM2 335OE. Percentage of miRNAs as in Figure 4.2 on a log scale. miR-335 and miR-335* are highlighted.

compared with LM2 Control ($\text{FDR} \leq 0.05$), however the majority of these miRNA are rare components of the miRNA profile ($\text{average\%} < 0.1$). Despite ~120 fold increased expression of miR-335 and miR-335* in LM2 335OE cells, both mature miRNAs processed from this precursor remain minor components of the AGO-miRNA profile representing only 0.42% and 0.045% of the total miRNA population respectively.

The asymmetrical increase in miR-335 and miR-335* representation in AGO-miRNA HITS-CLIP libraries raised the possibility that miR-335* is the dominant guide strand in LM2 335OE cells. Given that the selected strand determines the functional specificity of the miRISC complex, this would have profound implications for the targets regulated by pre-miR-335. The stem-loop structure of human miR-335 is shown in Figure 4.4. Functional miRNAs have been found to have lower thermodynamic stability at the 5' end, leading to more efficient loading into RISC (Khvorova et al., 2003). The 5' end of miR-335 contains multiple G-C base pairs, while the miR-335* 5' end is comprised of a stretch of five A-T base pairs, suggesting that miR-335* 5' end would be relatively less stable. Consistent with miR-335* being the guide strand, it is detected at significantly higher levels than miR-335 in MDA-MB-231 and all LM2 cell lines used in this study indicating that processing of exogenous pre-miR-335 is not responsible for the observed AGO-miRNA asymmetry (Figure 4.5).

TABLE 4.2: Analysis of the AGO-miRNA profiles of LM2 Control and LM2 335OE. Percentage of AGO-miRNA CLIP tags is shown as the average across the tree libraries for each data set, \log_2 fold change, p-value and FDR calculated using EdgeR.

miRNA	% LM2 Control	% LM2 335OE	Log2 Fold Change	p-value	FDR
miR-335*	0.007	0.419	6.647	7.61E-15	5.29E-12
miR-548o	0.134	0.009	-3.155	1.03E-05	0.0024
miR-3182	0.022	0.001	-4.317	1.03E-04	0.0179
miR-665	0.046	0.003	-2.936	1.56E-04	0.0217
miR-574-5p	0.643	0.132	-1.724	2.54E-04	0.0295
miR-519d	0.025	0.001	-3.966	3.34E-04	0.0332
miR-663	0.015	0.000	-5.080	3.82E-04	0.0332
<hr/>					
miR-335	0.005	0.045	4.007	0.003	0.119

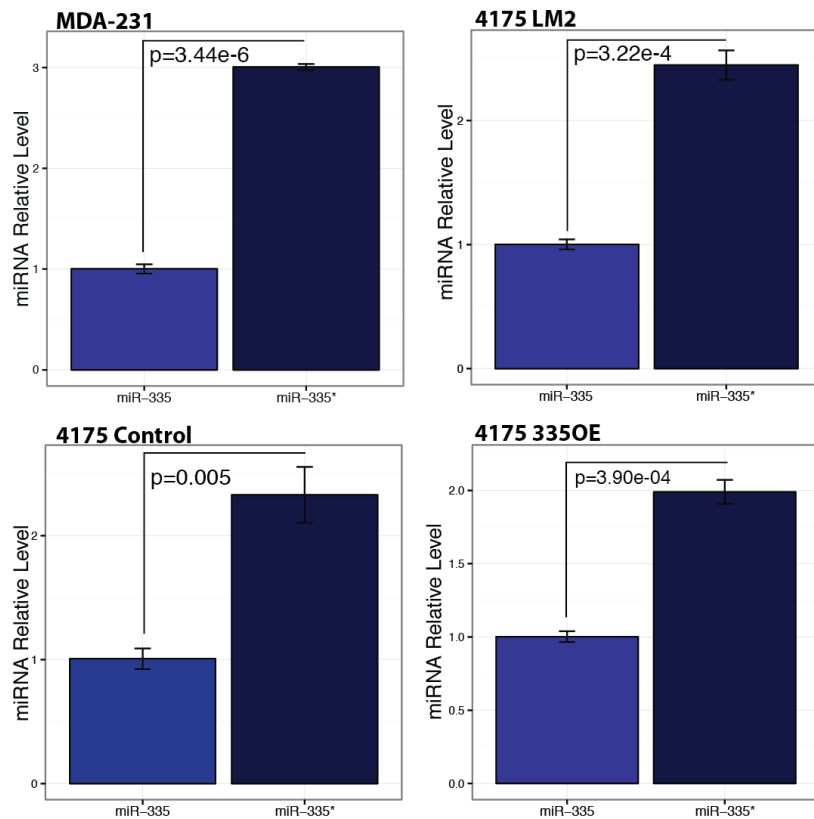


FIGURE 4.5: miR-335* is consistently detected at higher levels than miR-335 in MDA-231 and lung metastatic sub-lines. Q-PCR for miR-335 and miR-335* were carried out in triplicate for each cell line indicated with expression normalized to Snord44. Error bars = s.e.m, p-values calculated by t-test.

Global characterization of AGO-mRNA binding in LM2 Control and LM2 335OE

Uniquely mapped mRNA libraries were clustered independently for LM2 Control and LM2 335OE, and the resulting robust, reproducible clusters annotated ($BC \geq 2$ tags ≥ 5). As previously described, clusters mapping to chrY, RNA genes and pre-miRNA loci were removed before genomic distribution was assessed, resulting in 21,403 LM2 335OE and 31,359 LM2 Control clusters. The genomic localization of both LM2 Control and 335OE clusters were highly similar, with the largest proportion of clusters mapping to 3'UTRs plus 10kb immediately downstream of annotated gene ends (44% and 44.1%, control and 335OE respectively), and consistent with the previously described distribution of AGO-mRNA binding in 4175LM2 (see Figure 3.7). The majority of the remaining LM2 Control and 335OE clusters were distributed between coding sequence (30.5% and 27% respectively), introns (13.8% and 17.9% respectively), with a small degree of binding occurring in 5'UTR and deep intergenic regions (Figure 4.6).

To assess reproducibility between the AGO-mRNA binding sites in LM2 Control and LM2 335OE cells, all unique tags were pooled and clustered. These clusters were filtered to retain those with a minimum of five unique tags originating from two replicate libraries of either LM2 Control or 335OE, and by removing clusters mapping to chrY, RNA genes and pre-miRNA loci (39,109 AGO binding sites). The number of tags per cluster originating from each cell line was highly reproducible (Figure 4.7,

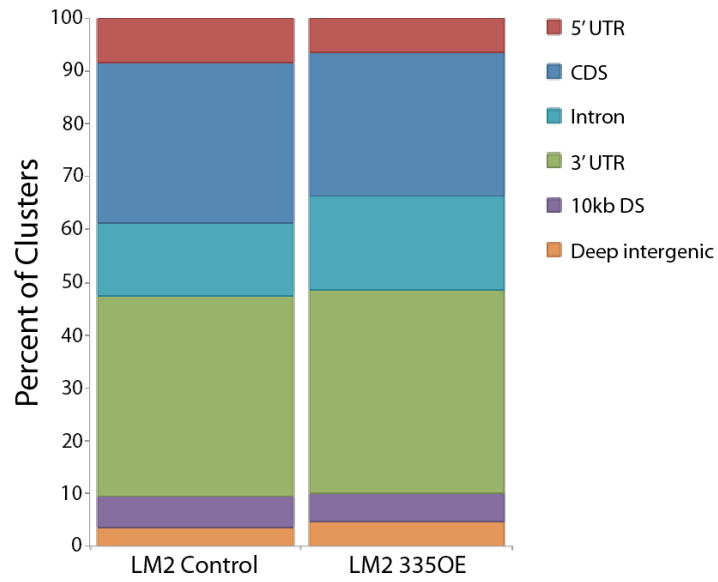


FIGURE 4.6: Genomic distribution of robust and reproducible AGO-mRNA clusters in LM2 Control and LM2 335OE cells. The three independently derived AGO-mRNA libraries for each cell line were pooled and clustered with a minimum overlap of 3nt. Clusters were removed that did not meet the minimum criteria of $BC \geq 2$ tags ≥ 5 , or that mapped to chrY, RNAGenes and pre-miRNA loci. Clusters were annotated to give genomic location.

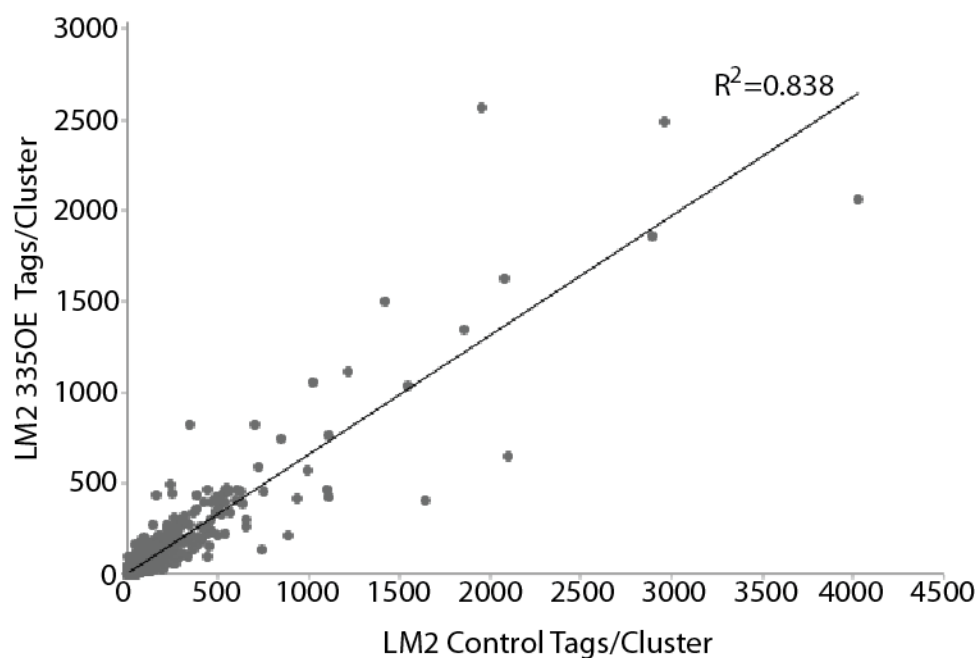


FIGURE 4.7: AGO-mRNA binding is highly reproducible between LM2 Control and LM2 335OE cells. All unique AGO-mRNA reads from LM2 Control and 335OE cells were pooled together and clustered, retaining only those clusters with $BC \geq 2$ and $tags \geq 5$ in either cell line. The scatter plot of CLIP tag per cluster originating from Control and LM2 cells shows a high degree of reproducibility in AGO-mRNA binding between cells lines ($R^2=0.838$). The overexpression of pri-miR-335 in LM2 335OE does not lead to a global re-assortment of AGO on target mRNAs.

$R^2=0.838$), suggesting that overexpression of pri-miR-335 does not globally perturb the AGO-mRNA profile of LM2 cells.

Identification of miR-335/ dependent de novo AGO-mRNA clusters*

The basic principles used to identify miR-335 or miR-335* dependent AGO binding events from AGO HITS-CLIP data are in outlined in Figure 4.8. First, AGO-mRNA clusters were identified that showed significant enrichment in AGO-mRNA binding LM2 335OE cells (dark blue) when compared to the non-targeting hairpin expressing LM2 Control cells (light blue) as previously described for differential AGO-mRNA analysis in Chapter 3. miR-335/*-dependent binding *de novo* binding events were then screened for the canonical miR-335 and/or miR-335* seed sequence (CTCTTG and TGAAAA) as evidence of direct miR-335/*-mediated AGO *de novo* binding.

A total of 1,283 $BC \geq 2$ $tag \geq 5$ clusters were identified with significant enrichment of AGO-binding in the presence of pri-miR-335 (\log_2 fold change 335OE/Control ≥ 1 , $p\text{-value} \leq 0.01$). 368 clusters (28.7%) contained a minimum 6-mer sequence match to miR-335 or miR-335* within position 1-8, while 210 (16.4%) contained a canonical seed match corresponding to position 2-7. Of those clusters with a canonical seed match, 71 contained seed matches to miR-335 and 139 to miR-335* consistent with the relatively higher abundance of miR-335* in the LM2 335OE AGO-miRNA profile. Restricting the location of clusters to the 3UTR and 10kb

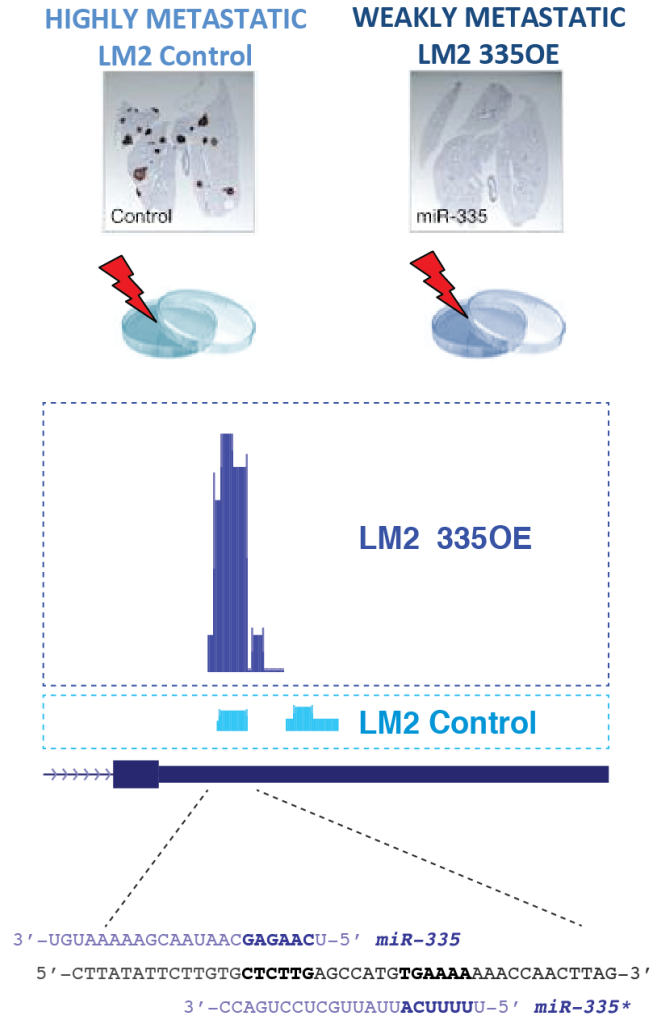


FIGURE 4.8: Experimental schematic of miR335/* dependent AGO-binding event identification. Triplicate AGO HITS-CLIP unique mRNA tags from 4175LM2 cells stably expressing either a control hairpin (LM2 Control, highly metastatic, light blue) or pri-miR-335 (LM2 335OE, weakly metastatic, dark blue) were mapped to the human genome (hg18). Sites of robust, reproducible binding were identified as either as clusters derived from pooled unique reads from both cell lines, or as significant peaks in the LM2 335OE data set (both requiring a minimal $BC \geq 2$ and $tag \geq 5$). Tag numbers from each cell line in the AGO-mRNA footprint region were counted and those AGO binding sites showing enrichment in the LM2 335OE libraries interrogated for the presence of miR-335 or miR-335* seed sequences as evidence of direct, *de novo* miR-335/* regulation. Lung section images from (Tavazoie et al., 2008).

downstream region left a total 107 miR-335/* canonical seed match containing, pri-miR-335-dependent clusters. Given that the definition of a cluster can encompass relatively large stretches of transcripts with multiple peaks, *de novo* miR-335/* AGO-mRNA binding was assessed in the -30/+32nt footprint centered on the single largest peak of each cluster (Chi et al., 2009). Seed search in this narrower sequence space returned 106 miR-335* and 66 miR-335 6mer seed matches to positions 1-8 (13.4% of *de novo* AGO-miR335 clusters). Of these peaks, 64 and 35 contained canonical 2-7 seed matches to miR-335* and miR-335 respectively, with 44 canonical seed matches occurring in 39 3'UTR and 10kb downstream clusters (Table 4.3).

Detailed analysis of AGO HITS-CLIP data sets have revealed alternative rules for miRNA:mRNA target recognition involving the inclusion of position 6 nucleation bulges (Chi et al., 2012) and various “seed-like” sequences (Loeb et al., 2012). Interestingly, G:U wobbles can be tolerated in various positions within the *lsy-6::cog-1* targeting site, which like miR-335* seed pairing contains a stretch of four A-U base pairs (TACAAAA – *lsy-6*, ATGAAAA - miR-335*) (Didiano and Hobert, 2006). Taking these potential alternative seed pairing sequences into consideration, a further 133 potential seed matches for miR-335* were identified in the 62nt AGO footprint of those *de novo* clusters with no evidence of a canonical 6nt seed (Figure 4.9). The ability of miR-335* to regulate these non-canonical seed-matched transcripts is yet to be determined. MEME analysis of the remaining non-miR-335/* *de novo*

TABLE 4.3: LM2 335OE 3'UTR clusters containing canonical seed matches in the 62nt AGO footprint. *De novo* mR-335/* 3'UTR clusters whose largest peak (-30/+32nt) contain a canonical seed match to either miR-335 or miR-335*. Unique mRNA counts and biological complexity are given for LM2 335OE and LM2 Control cells with \log_2 fold change in binding events (normalized to read depth per cell line), associated p-value and FDR (Benjamini-Hochberg correction). The genomic coordinates of the AGO cluster region and the seed match are given for reference.

GENE	miRNA	Seed Match	335OE		Control		Log2 FC			p-value	FDR	Location	CHRM	Start	End
			Tags	BC	Tags	BC	335OE/Cont								
DBT	miR-335*	TGAAAA	9	2	0	0	3.83	0.0010	0.0490	3'UTR	chr1	100434187	100434237		
KDSR	miR-335	CTCTTG	8	3	0	0	3.68	0.0022	0.0826	3'UTR	chr18	59148339	59148403		
KDSR	miR-335	CTCTTG	8	3	0	0	3.68	0.0022	0.0826	3'UTR	chr18	59148339	59148403		
SEL1L	miR-335*	TGAAAA	7	2	0	0	3.51	0.0048	0.1298	3'UTR	chr14	81012555	81012597		
SEL1L	miR-335	CTCTTG	7	2	0	0	3.51	0.0048	0.1298	3'UTR	chr14	81012555	81012597		
CDH11	miR-335	CTCTTG	6	3	0	0	3.32	0.0103	0.1817	3'UTR	chr16	63538440	63538508		
CDC7	miR-335	CTCTTG	6	2	0	0	3.32	0.0103	0.1838	3'UTR	chr1	91763846	91763876		
MCM3APAS	miR-335*	TGAAAA	6	2	0	0	3.32	0.0103	0.1833	3'UTR	chr21	46495888	46495991		
PLAGL2	miR-335	CTCTTG	28	3	5	2	2.78	7.49E-07	0.0001	3'UTR	chr20	30244118	30244205		
AAK1	miR-335*	TGAAAA	8	2	1	1	2.68	0.0104	0.1751	3'UTR	chr2	69542097	69542148		
PBX3	miR-335*	TGAAAA	8	3	1	1	2.68	0.0104	0.1746	3'UTR	chr9	127769298	127769369		
NEDD4L	miR-335	CTCTTG	11	2	2	1	2.51	0.0030	0.1001	3'UTR	chr18	54215785	54215878		
PANX1	miR-335*	TGAAAA	11	3	2	2	2.51	0.0030	0.0998	3'UTR	chr11	93554290	93554423		
SDAD1	miR-335*	TGAAAA	10	2	2	1	2.39	0.0057	0.1375	3'UTR	chr4	77090781	77090857		
SFRS2	miR-335*	TGAAAA	9	3	2	1	2.25	0.0110	0.1778	3'UTR	chr17	72242674	72242787		
CETN3	miR-335*	TGAAAA	9	3	2	2	2.25	0.0110	0.1768	3'UTR	chr5	89725520	89725590		
PTPN3	miR-335	CTCTTG	12	3	3	3	2.21	0.0050	0.1302	3'UTR	chr9	111180764	111180990		
CTR9	miR-335*	TGAAAA	12	3	3	3	2.21	0.0050	0.1308	3'UTR	chr11	10757431	10757522		
SMC4	miR-335*	TGAAAA	19	3	6	3	2.03	0.0007	0.0400	3'UTR	chr3	161612959	161613087		
SLC4A7	miR-335*	TGAAAA	24	3	8	3	1.99	0.0002	0.0150	3'UTR	chr3	27390911	27391018		
GPR180	miR-335*	TGAAAA	15	3	6	2	1.70	0.0094	0.1904	3'UTR	chr13	94077653	94077706		
TMED5	miR-335*	TGAAAA	17	3	7	3	1.68	0.0047	0.1398	3'UTR	chr1	93392602	93392740		
TOP2A	miR-335	CTCTTG	27	3	13	3	1.51	0.0014	0.0643	3'UTR	chr17	35798579	35798692		
POLR3F	miR-335*	TGAAAA	21	3	10	3	1.51	0.0041	0.1235	3'UTR	chr20	18412263	18412340		

GENE	miRNA	Seed Match	335OE		Control		Log2 FC		FDR	Location	CHRM	Start	End
			Tags	BC	Tags	BC	335OE/Cont	p-value					
FAM105B	miR-335*	TGAAAA	22	3	11	3	1.45	0.0052	0.1341	3'UTR	chr5	14751036	14751131
ABCF1	miR-335	CTCTTG	20	3	10	3	1.44	0.0065	0.1512	3'UTR	chr6	30666889	30667074
ACSL4	miR-335*	TGAAAA	18	3	9	3	1.44	0.0127	0.1954	CDS/3'UTR	chrX	108773847	108774036
SMAD5	miR-335*	TGAAAA	32	3	17	3	1.39	0.0009	0.0477	3'UTR	chr5	135541063	135541291
ANP32E	miR-335	CTCTTG	41	3	22	3	1.38	0.0002	0.0167	3'UTR	chr1	148458998	148459325
TMEM49	miR-335*	TGAAAA	26	3	14	3	1.36	0.0028	0.0968	CDS/3'UTR	chr17	55271858	55272050
PDIA6	miR-335*	TGAAAA	29	3	16	3	1.33	0.0026	0.0966	3'UTR	chr2	10841607	10841810
LPAR1	miR-335*	TGAAAA	43	3	25	3	1.27	0.0004	0.0259	3'UTR	chr9	112677209	112677394
LPAR1	miR-335	CTCTTG	43	3	25	3	1.27	0.0004	0.0259	3'UTR	chr9	112677209	112677394
ATP13A3	miR-335	CTCTTG	27	3	16	3	1.23	0.0058	0.1390	3'UTR	chr3	195607846	195608126
ZBTB34	miR-335*	TGAAAA	22	3	13	3	1.23	0.0118	0.1841	3'UTR	chr9	128687426	128687576
ZBTB34	miR-335	CTCTTG	22	3	13	3	1.23	0.0118	0.1841	3'UTR	chr9	128687426	128687576
MAT2B	miR-335*	TGAAAA	40	3	24	3	1.23	0.0007	0.0399	3'UTR	chr5	162877977	162878152
MAT2B	miR-335	CTCTTG	40	3	24	3	1.23	0.0007	0.0399	3'UTR	chr5	162877977	162878152
GLT8D3	miR-335*	TGAAAA	29	3	18	3	1.17	0.0055	0.1345	3'UTR	chr12	40766833	40767085
CEP350	miR-335	CTCTTG	25	3	16	3	1.12	0.0127	0.1952	3'UTR	chr1	178348537	178348783
MAPIB	miR-335*	TGAAAA	124	3	81	3	1.12	3.74E-08	0.0000	3'UTR	chr5	71540822	71541150
MAPIB	miR-335	CTCTTG	124	3	81	3	1.12	3.74E-08	0.0000	3'UTR	chr5	71540822	71541150
HN1	miR-335	CTCTTG	86	3	59	3	1.05	1.58E-05	0.0021	CDS/3'UTR	chr17	70643666	70643957
TMEM41B	miR-335*	TGAAAA	41	3	28	3	1.05	0.0025	0.0918	3'UTR	chr11	9261244	9261449

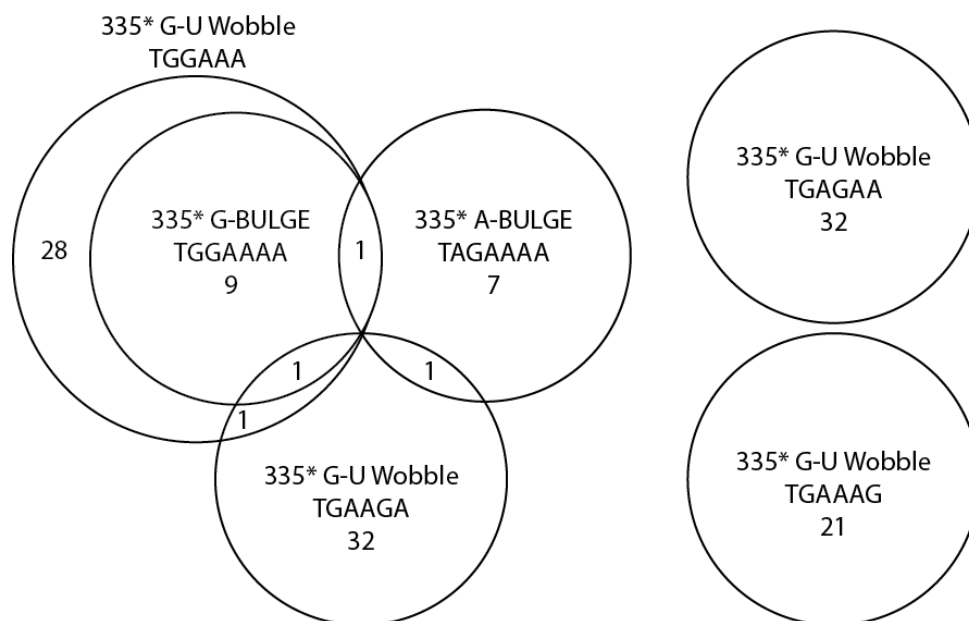


Figure 4.9: Non-canonical miR-335* seed sites in the AGO 62nt footprint in LM2 335OE enriched AGO clusters. AGO footprint regions enriched in LM2 335OE cells were searched for potential non-canonical miR-335 and miR-335* seed sites with bulged nucleotides at position 6 (Chi et al., 2012), and G:U wobbles in the 4nt A-U pairing region of miR-335* (Didiano and Hobert, 2006). 133 additional binding sites were found that contained a non-canonical miR-335* seed-like sequence, no bulged sites were found for miR-335.

clusters identified two significant motifs, which did not match the seed sequences of any miRNA in the HITS-CLIP profile, or the G-rich motifs associated with miRNA-independent AGO binding sites identified in mouse embryonic stem cells (Leung et al., 2011).

Identification of miR-335/ dependent de novo AGO-mRNA significant peaks*

Unique mRNA tags from LM2 335OE were used to identify significant peaks of AGO binding. 9,183 significant peaks were identified with $p \leq 0.01$, 7305 of which contained five or more tags from two or more replicate LM2 335OE libraries ($BC \geq 2$ tags ≥ 5). The number of LM2 Control AGO-mRNA tags overlapping the ± 30 nt region centered on the LM2 335OE significant peak was calculated and difference analysis carried out as for clusters. 253 LM2 335OE significant peaks showed significant enrichment of AGO-mRNA binding (\log_2 fold change ≥ 1 , $p < 0.01$), with 82 containing a minimal 6-mer match to either miR-335 or miR-335* between positions 1-8. Only 15 peaks contained canonical position 2-7 seed matches, of which 9 were located in 3'UTR regions. Table 4.4 shows all canonical seed-matched significant peaks in 3'UTRs and coding sequence. The low number of seed containing significant peaks identified as enriched in LM2 335OE is unsurprising considering the small proportion of the miRNA profile represented by miR-335 and -335*, and the generally low tag numbers found in *de novo* miR-335/*-dependent and seed containing clusters.

Table 4.4: LM2 335OE enriched significant peaks in CDS and 3'UTRs with canonical miR-335/* seed sequences. Significant peaks were defined in LM2 335OE cells ($BC \geq 2$ tags ≥ 5), and overlapping tags from LM2 Control cells counted in a ± 30 nt AGO footprint centered on the 335OE significant peak coordinate. Difference analysis was carried out as described for cluster analysis. 3'UTRs and CDS significant peaks that contain a canonical miR335 or miR335* seed site and have significant binding enrichment in LM2 335OE cells are shown.

GENE	miRNA	Seed Match	335OE		Control		Log2FC	p-value	FDR	Location	CHRM	Start	End
			Tags	BC	Tags	BC							
KDSR	miR-335	1-8	8	3	0	0	3.27	0.0094	0.1506	3'UTR	chr18	59148338	59148398
CALU	miR-335	1-8	18	3	2	1	2.77	0.0004	0.0144	3'UTR	chr7	128198240	128198300
GHITM	miR-335*	2-7	11	3	1	1	2.69	0.0058	0.1146	CDS	chr10	85893789	85893849
PLAGL2	miR-335	2-7	28	3	5	2	2.38	2.93E-05	0.0020	3'UTR	chr20	30244117	30244177
SERPINE1	miR-335	2-8	43	3	10	3	2.10	2.07E-06	0.0002	3'UTR	chr7	100568166	100568226
DKK1	miR-335*	2-8	31	3	7	3	2.10	4.83E-05	0.0029	3'UTR	chr10	53746593	53746653
DST	miR-335*	2-7	17	3	4	3	1.95	0.0054	0.1102	CDS	chr6	56674647	56674707
ANP32E	miR-335	1-7	34	3	9	3	1.91	6.95E-05	0.0040	3'UTR	chr1	148459186	148459246
CEP350	miR-335	1-8	22	3	8	3	1.46	0.0078	0.1386	3'UTR	chr1	178348632	178348692
DKK1	miR-335	2-7	29	3	11	3	1.42	0.0030	0.0719	3'UTR	chr10	53746972	53747032
MAP1B	miR-335	1-7	95	3	49	3	1.04	2.13E-05	0.0015	3'UTR	chr5	71540874	71540934

Initial selection of miR-335/ targets for validation*

The list of candidate miR-335/* enriched targets was refined by comparing to RNAseq analysis of transcript abundance in LM2 Control and 335OE cells. Single-end, 50nt polyA-selected single replicate RNA libraries were prepared from LM2 Control and LM2 335OE cells harvested at the same time as those used for HITS-CLIP. The resulting high-throughput sequencing data was filtered based on quality score then aligned to the human genome (hg18), giving ~20million reads for LM2 Control and ~21million reads for LM2 335OE. Gene expression levels (RPKM) were obtained by aligning reads to a non-redundant set of constitutive exons, and difference in expression calculated using tools developed by Chaolin Zhang in the Darnell laboratory for use within the Galaxy bioinformatics platform.

RPKM values per gene from LM2 Control and LM2 335OE cells are plotted in Figure 4.10, with the previously identified miR-335 downregulated gene set (Tavazoie et al., 2008) highlighted in blue as a positive control. A high degree of correlation in transcript abundance ($R^2=0.927$) was observed between LM2 335OE and LM2 Control indicating very few transcripts with altered steady state levels. Filtering the LM2 335OE enriched clusters by \log_2 fold change of 335OE/Control RNAseq reads (≤ 0.2 , $p\text{-value} \leq 0.01$) returned 9 clusters, of which 8 clusters contained canonical miR-335* seed matches, one contained a miR-335 seed while three clusters contained both seed sequences (Table 4.5).

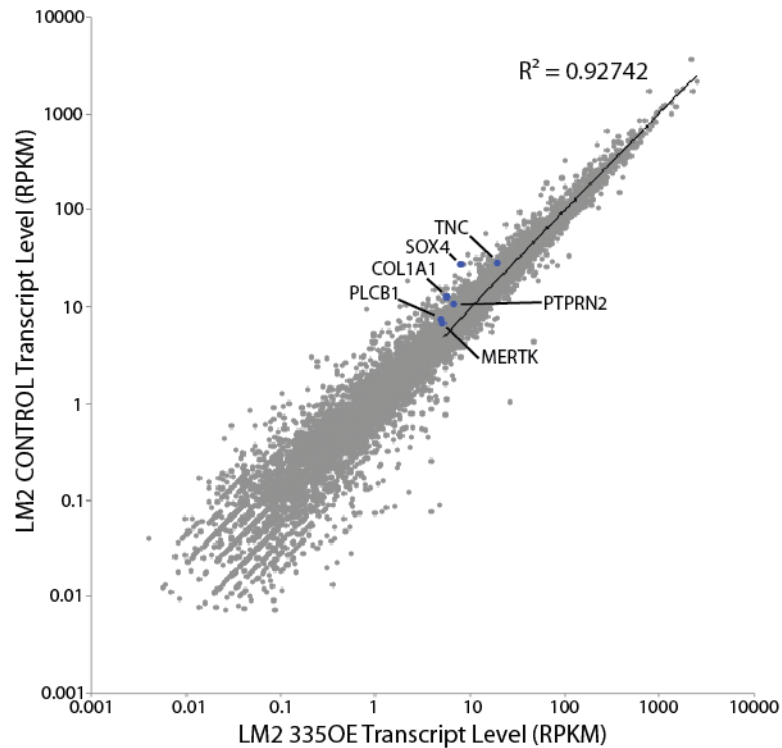


FIGURE 4.10: RNAseq analysis of steady state transcript levels in LM2 Control and LM2 335OE cells. Scatter plot of RPKM values per gene obtained from high-throughput polyA-selected RNAseq. Genes highlighted in blue comprise the six miR-335 downregulated genes previously identified in LM2 335OE cells (Tavazoie et al., 2008), with all six showing derepression in LM2 Control cells.

Table 4.5: miR-335/* canonical seed containing LM2 335OE-enriched AGO clusters downregulated in LM 335OE. Clusters shown are BC ≥ 2 tag ≥ 5 , fold change ≥ 1 with associated p-value <0.01 , filtered by RNAseq FC ≤ 0.2 and associated p-value ≤ 0.01 .

GENE	miRNA	Seed Match	335OE			Control			CLIP Log2 FC			RNAseq Log2 FC		
			Tags	BC	Tags	BC	Tags	BC	335OE/Cont	p-value	FDR	335OE/Control	p-value	FDR
SEL1L	miR-335*	2-7	7	2	0	0	0	0	3.512	0.0048	0.1298	-0.0911	0.0025	0.0100
SEL1L	miR-335	2-7	7	2	0	0	0	0	3.512	0.0048	0.1298	-0.0911	0.0025	0.0100
CETN3	miR-335*	1-7	9	3	2	2	2	2	2.249	0.0110	0.1768	-0.2759	0.0004	0.0021
TOP2A	miR-335	1-8	27	3	13	3	3	3	1.512	0.0014	0.0643	-0.0829	3.68E-07	3.12E-06
FAM105B	miR-335*	2-8	22	3	11	3	3	3	1.450	0.0052	0.1341	-0.2883	1.43E-07	1.28E-06
ACSL4	miR-335*	2-7	18	3	9	3	3	3	1.438	0.0127	0.1954	0.1502	6.13E-13	9.85E-12
SMAD5	miR-335*	2-7	32	3	17	3	3	3	1.386	0.0009	0.0477	-0.1620	8.46E-08	7.89E-07
MAT2B	miR-335*	1-8	40	3	24	3	3	3	1.225	0.0007	0.0399	-0.2273	9.80E-08	9.01E-07
MAT2B	miR-335	1-8	40	3	24	3	3	3	1.225	0.0007	0.0399	-0.2273	9.80E-08	9.01E-07
MAP1B	miR-335*	1-7	124	3	81	3	3	3	1.120	0.0030	3.74E-08	-0.0803	3.88E-05	0.0002
MAP1B	miR-335	1-7	124	3	81	3	3	3	1.120	2.13E-05	3.74E-08	-0.0803	3.88E-05	0.0002
TMEM41B	miR-335*	2-7	41	3	28	3	3	3	1.046	0.0025	0.0918	-0.1842	6.24E-06	4.40E-05

Validation of centrin EF-hand protein 3 (CETN3) as a miR-335 target*

CETN3 was selected for validation based on the high relative enrichment of CLIP tags and fold change in abundance compared with the other candidates. Additionally, the *de novo* AGO binding site identified by differential analysis was the only AGO binding site in this relatively short 3'UTR (~350nt, Figure 4.11A), and contained extensive complementarity to miR-335*; a canonical miR-335* 1-7nt seed match with an additional G-U wobble base-pair at position 8 (Figure 4.11B). Endogenous CETN3 transcript levels were assayed by Q-PCR in MDA-231 parental cells (miR-335/* expressing), 4175 LM2 (miR-335/* negative) and LM2 335OE. CETN3 steady state mRNA levels were significantly higher in 4175LM2 cells suggesting derepression of the transcript (Figure 4.12A). To determine whether this derepression was a direct consequence of decreased levels of miR-335*, 4175LM2 cells were transiently transfected with miR-335* mimic. This resulted in significant down regulation of endogenous CETN3, while no effect on transcript level was seen in cells transfected with miR-335 mimic or a negative control mimic that is non-targeting in the human genome (p=0.006, Figure 4.12B).

To determine whether the miR-335* seed site identified within the CETN3 3'UTR *de novo* AGO cluster mediated miR-335*-dependent transcript repression, full length WT CETN3 3'UTR was cloned into a dual luciferase reporter construct (psiCheck2). A mutant construct was made altering the miR-335* seed site at two positions, from TGAAAA to TGATAT. Transient transfection of these constructs into 4175LM2 was

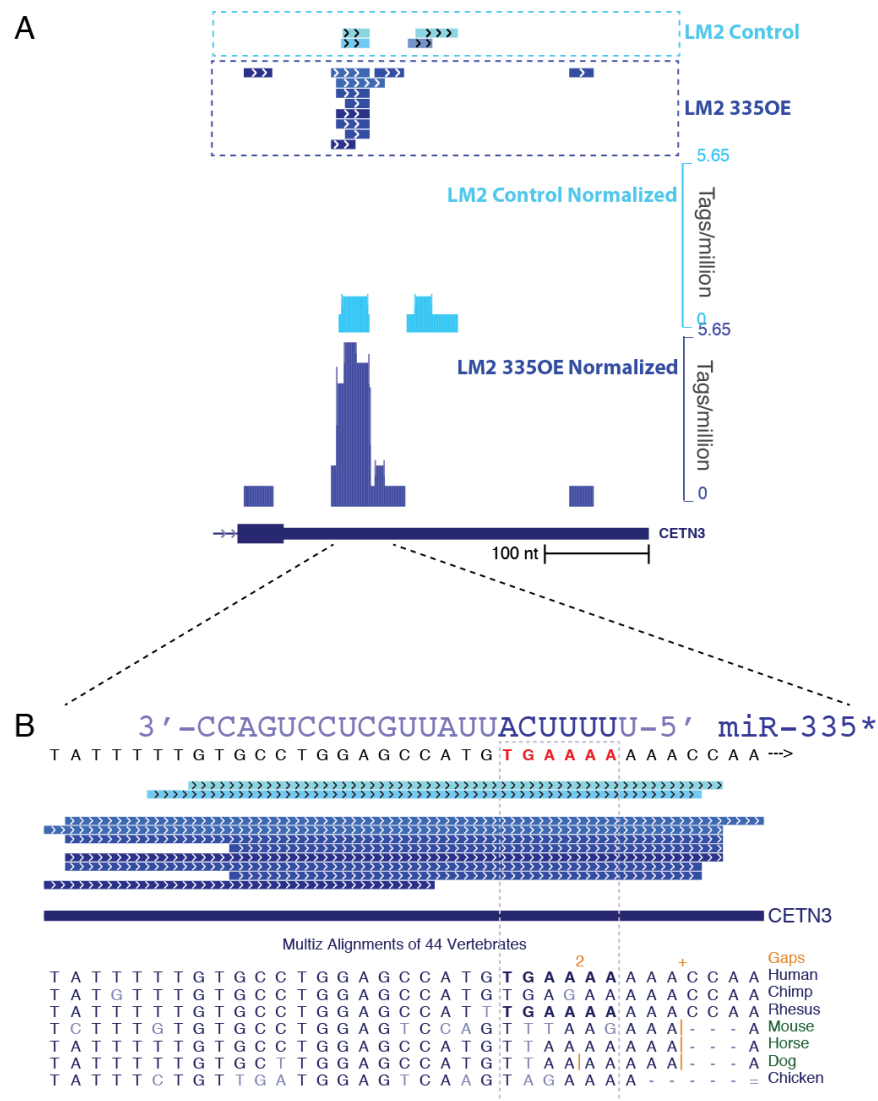


FIGURE 4.11: Differential AGO binding identifies CETN3 as a miR-335* target. **A:** Binding map of AGO:mRNA unique tags (top panel) and normalized reads (bottom panel) from LM2 Control (light blue) and LM2 335OE cells (dark blue) showing increased binding in a discrete site on the 3'UTR of CETN3 in LM2 335OE. **B:** The footprint of the differential AGO binding site contains a non-conserved miR-335* seed match.

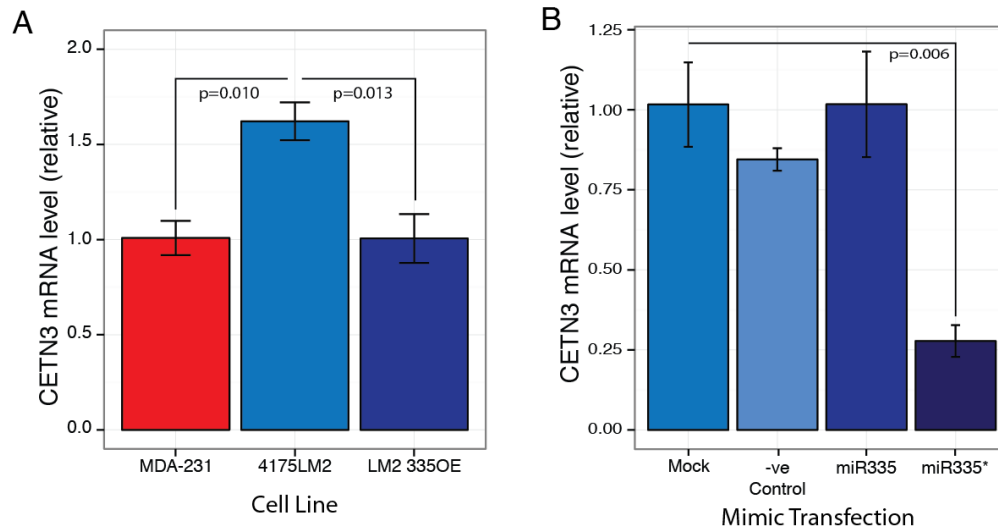


FIGURE 4.12: CETN3 mRNA level is repressed by miR-335*. **A:** CETN3 mRNA levels are significantly higher in 4175 LM2 cells (miR-335/* negative) than in MDA-231 cells (miR-335/* positive). CETN3 mRNA levels return the baseline level seen in MDA-231 cells in LM2 335OE cells. **B:** CETN3 mRNA level is negatively regulated by transient transfection of 50nM miR-335* mimic into 4175LM2 cells, but not miR-335 mimic. Mock is transfection reagent alone, negative control miRNA mimic is non-targeting in human. n=3 cell line, n=4 transfection Q-PCR, expression levels normalized to TBP, error bars = s.e.m, p-values calculated by t-test.

carried out in the presence of either negative control miRNA mimic or miR-335* mimic. WT CETN3 3'UTR luciferase reporter activity was significantly inhibited in the presence of miR-335* mimic ($p=4.6e^{-5}$), while the mutant CETN3 3'UTR was unaffected (Figure 4.13A). This miR-335* seed-dependent regulation of the CETN3 3'UTR was verified in an independent cell line (HEK 293T, Figure 4.13B).

The ability of endogenous miR-335* to repress the CETN3 3'UTR was assayed in MDA-231 cells. Initial transfections of 50ng WT and mutant CETN3 3'UTR reporter constructs into MDA-231 and 4175LM2 indicated increased basal reporter activity of WT CETN3 3'UTR in 4175LM2 compared to the MDA-231 parental line, however transfections with the mutant construct failed to reveal derepression in MDA-231 (data not shown). In attempt to introduce reporter construct at more stoichiometric concentrations, 5ng of plasmid was transfected into cells with high replicate numbers ($n=10$). Expression of WT CETN3 was significantly repressed when compared to mutant CETN3 3'UTR in MDA-231 cells where miR-335* is expressed. No significant difference between WT and mutant reporter activity was seen in 4175LM2 cells where miR-335* expression is extremely low (Figure 4.14).

To investigate whether upregulation of CETN3 in miR-335* expressing cells could enhance or restore invasive capability, stable MDA-231 and LM2 335OE cell lines were generated that overexpressed human

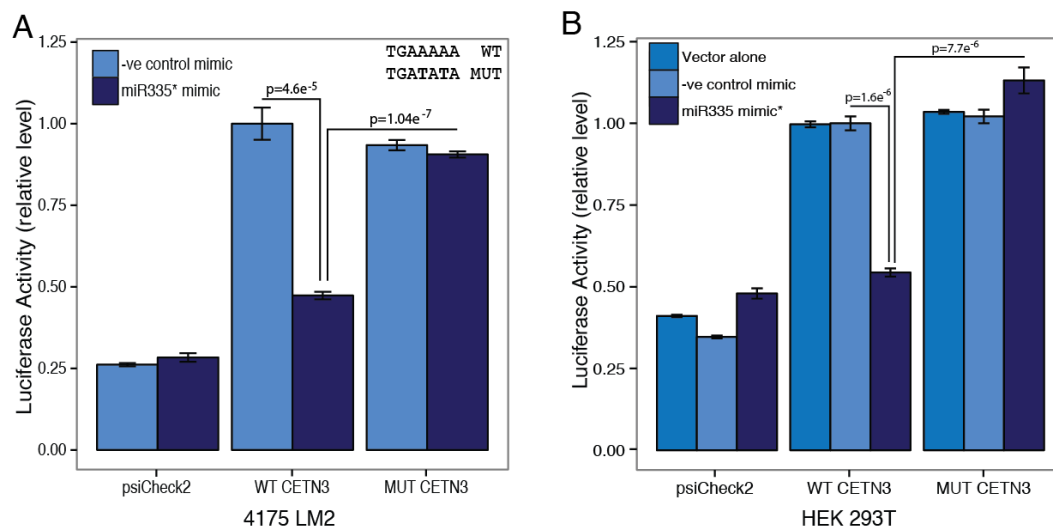


FIGURE 4.13: miR-335* represses the 3'UTR of CETN3. 3'UTR reporter assays were carried out in 4175LM2 cells (A) and verified in HEK-293T (B). 50ng of dual luciferase reporter constructs encoding the full length WT or mutant CETN3 3'UTR, or empty vector (psiCheck2) were transfected in the cell line indicated either alone, or with 50nM miRNA mimic as indicated. Luciferase activity was assayed 30h after transfection and values normalized to WT CETN3 transfected with negative mimic. Co-transfection of miR-335* mimic leads to significant repression of the luciferase activity of WT CETN3 3'UTR, which is abrogated by mutating the miR-335* seed site identified in the differentially bound AGO-footprint. n=4, error bars = s.e.m, p-values calculated by t-test.

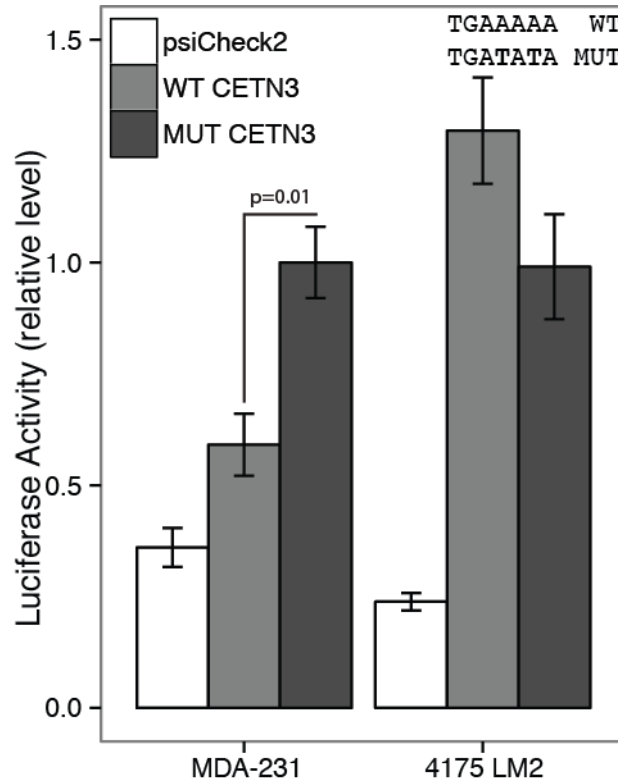


FIGURE 4.14: Endogenous miR-335* represses the CETN3 3'UTR. Mutation of the miR-335* seed site bound by AGO in CETN3 3'UTR leads to derepression of reporter constructs in MDA-231 cells where miR-335* is expressed but not in 4175LM2 cells. Cells were transfected with 5ng of psiCheck2 empty vector, full length WT CETN3 or MUT CENT3 3'UTR and assayed for luciferase activity at 30h. Values are shown normalized to MUT CETN3 in MDA-231, n=10, error bars = s.e.m, p-values calculated by t-test.

CETN3 cDNA. Overexpression was validated by Q-PCR (Figure 4.15A) and the effect of increased CETN3 expression on invasiveness assayed by matrigel invasion. CETN3 overexpression in LM2 335OE cells had no effect on invasion compared to empty vector control cells. Overexpression of CETN3 in MDA-231 lead to a significant increase in invasion ($p=0.0106$) compared to MDA-231 uninfected controls ($n=4$, shown in Figure 4.15B), however this result failed to repeat in independent experiments, therefore the role of CETN3 in tumor cell invasion is currently undetermined.

Clinical association of CETN3 and metastasis free survival

Given that low miR-335 expression is negatively correlated with metastasis free survival in breast cancer patients (Tavazoie et al., 2008), it would be anticipated that high levels of miR-335/*- regulated targets would negatively correlate with prognosis. To determine the relationship between CETN3 expression levels and human breast cancer metastasis two data sets were probed. The first, “Kaplan-Meier Plotter” (www.kmplot.com, 2012 release, (Gyorffy et al., 2010)) uses a background database of publicly available primary breast cancer GEO expression data and survival information ($n=2,978$). In this data set, high levels of CETN3 were significantly correlated with poor metastasis-free survival (Figure 4.16A, $p=0.00018$ using default settings). However, using a second data set derived from patients after taxane-anthacycline chemotherapy treatment (Hatzis et al., 2011) that is not included in the Kaplan-Meier Plotter data set showed an opposite correlation, with high levels of CETN3 predictive for good metastasis-free survival (Figure 4.16B, $p=0.035$ calculated using

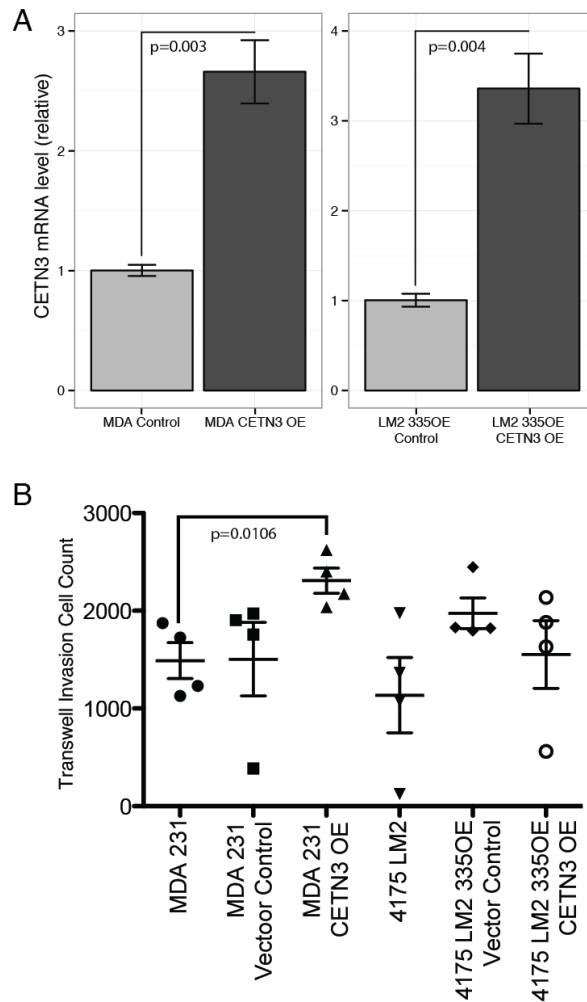


FIGURE 4.15: Investigating the role of CETN3 in invasion. Stable CETN3 overexpressing MDA-231 and LM2 335OE cell lines were generated via viral transduction with CETN3 cDNA. Empty pBabe-Puro vector was used as a control. **A:** Over expression of CETN3 was assayed by Q-PCR with RNA independently prepared from triplicate samples, as in Figure 4.12. **B:** Trans-well invasiveness of CETN3 OE cell lines was assessed by comparison to empty vector transduced cells and non-infected cells. Cells that had invaded through the matrigel inserts were counted 24h post seeding. Each cell line was tested in quadruplicate with cell triplicated cell counts taken per insert. No effect of CETN3 overexpression on invasion was seen in LM2 335OE cells, while over expression of CETN3 in MDA-231 cells lead to a significantly enhanced invasion (t-test) compared to MDA-231 uninfected controls but not to MDA-231 empty-vector stable cell line controls. This result did not repeat in a replicate experiment.

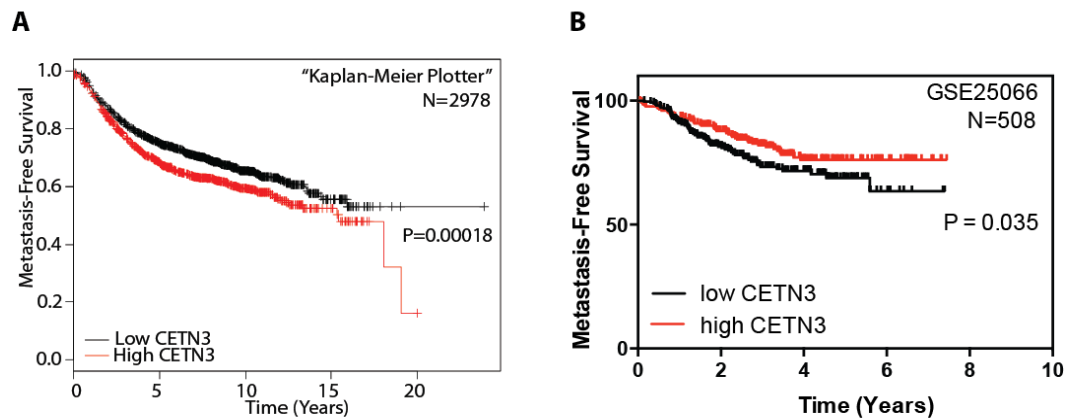


FIGURE 4.16: Clinical association of CETN3 expression level and metastasis-free survival. **A:** High levels of CETN3 correlate with shorter metastasis-free survival in breast cancer patients included in the Kaplan-Meier Plotter database (N=2978). CETN3 (209662_at) expression level classified based on median probe intensity with all settings at default. **B:** Low levels of CETN3 correlate with shorter metastasis-free survival in a 508 patient cohort whose gene expression profiles were generated after chemotherapy treatment.

the survdiff function in R by Hani Goodarzi, Tavazoie Lab). Whether this difference is attributable to chemotherapy treatment, and how CETN3 expression correlates with miR-335/* levels in these patient cohorts is unknown.

Further validation of predicted miR-335/ targets identified by AGO-HITS CLIP*

SEL1L (miR-335 and -335* seed), FAM105B (miR-335*) and SMAD5 (miR-335*) steady state transcript levels were assayed by Q-PCR in MDA-231, 4175 LM2 and LM2 335OE. These genes have significant enrichment in a miR-335/* seed containing AGO binding site in LM2 335OE cells (Figure 4.17A, $\log_2 FC \geq 1$, $p < 0.01$) and a predicted decrease in transcript level by RNAseq, although predicted expression level changes were small (Table 4.5). All three transcripts had increased steady state levels in 4175LM2 cells compared to pre-miR-335 expressing MDA-231, although this increase failed to reach significance for FAM105B. Transcript levels failed to show statistically significant repression in LM2 335OE cells when compared to 4175LM2 (Figure 4.17B).

Transient transfection of relevant miR-335/* mimics in LM2 cells also failed to repress these endogenous transcripts when compared to mock transfection controls (Figure 4.17C). Up regulation of both SEL1L and FAM105B was seen upon transfection with negative control mimic, which for FAM105B was significantly reversed by either miR-335 or miR-335* mimic transfection. This indicates both the potential for target-dependent general derepression of transcripts as a result of off-target

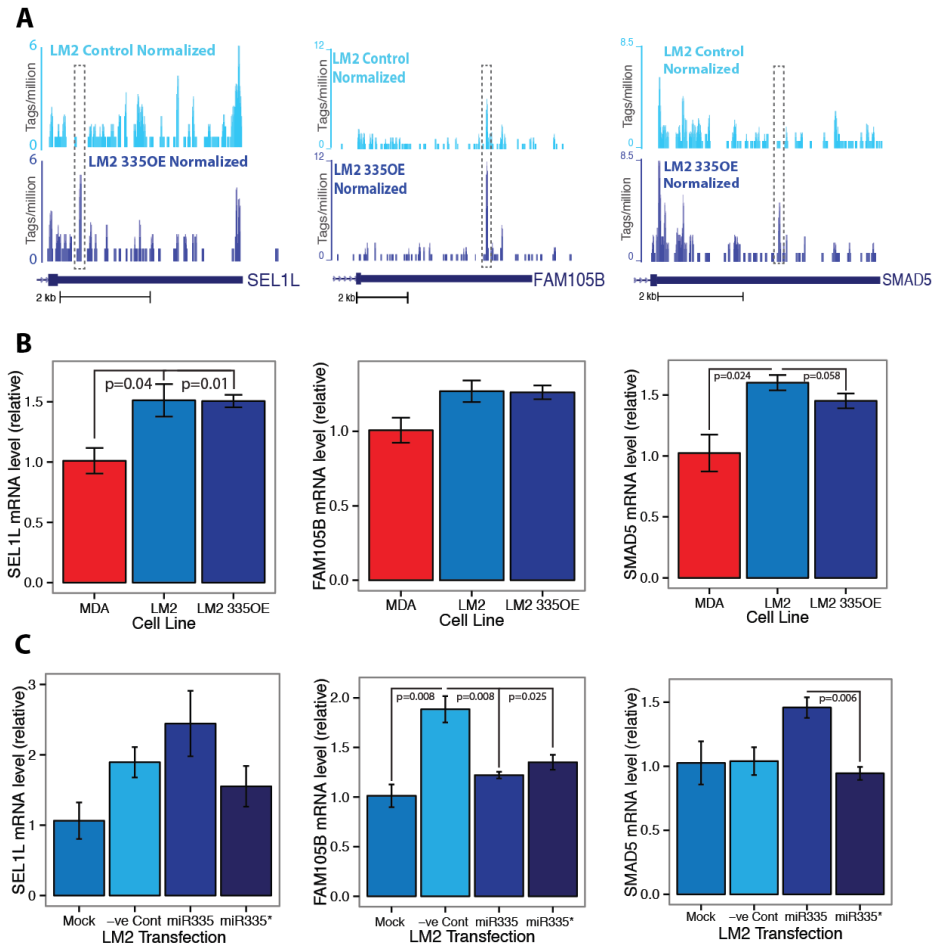


Figure 4.17: Steady state mRNA levels of predicted miR-335/* targets show weak or no regulation by endogenous or exogenous miR-335/*.

A: Extensive AGO binding is seen in the 3'UTRs of SEL1L, FAM105B and SMAD5 in both LM2 Control and LM2 335OE (normalized AGO binding). The pre-miR335 seen containing *de novo* AGO binding site is highlighted. **B:** Steady state levels of miR-335/* bound transcripts in MDA-231 (pre-miR-335 positive), 4175LM2 (pre-miR-335 negative) and LM2 335OE consistently show up regulation in 4175LM2 that is not reduced to MDA-231 levels in LM2 335OE. **C:** Transient transfection of 50nM miRNA mimics in LM2 cells fails to repress endogenous predicted target mRNAs when compared to mock transfection. For SEL1L and FAM105B, negative control mimic leads to target up-regulation, possibly through competition for AGO. The non-specific up regulation of FAM105B is reduced by miR-335 or miR335* mimic transfection, however miR-335 is not predicted to target FAM105 by HITS-CLIP. n=3 cell line, n=4 mimic transfection, expression levels normalized to TBP, error bars = s.e.m, p-values calculated by t-test.

effects of exogenous miRNA mimic transfection (Khan et al., 2009), and the potential for false positive result from reporter assays. FAM105B contains multiple miR-335 seed sequences within its 3'UTR that show no evidence of direct AGO binding by HITS-CLIP, suggesting that the miR-335 repressive effect is an artifact. As can be seen in Figure 4.17A, the *de novo* pre-miR-335 dependent AGO binding sites in these 3'UTRs are a relatively minor component of the total AGO-mRNA binding profile. The majority of AGO binding is common between LM2 Control and 335OE, suggesting that any additional anticipated negative regulation exerted by *de novo* miR-335/* binding sites in these transcripts is negligible or below the level detection by our assays in the background of high levels of shared repressive binding. Alternatively, these binding events may reflect a degree of bioinformatically-generated noise due to the low tag numbers associated with *de novo* miR-335/*-dependent AGO binding sites. Whether miR-335* can regulate SEL1L, FAM105B and SMAD5 protein level is yet to be determined.

miR-335/* regulation associated with *de novo* pri-miR-335 AGO binding is detectable in other transcripts that have a shared background of AGO binding. For example, enriched *de novo* AGO binding in twisted gastrulation homolog 1, TWSG1 (Log2 FC CLIP cluster=1.97, p-value=0.016), centered on a miR-335 seed with extensive additional complementarity (Figure 4.18A) correlates with steady state expression level changes with respect to endogenous or exogenous miR-335 (Figure 4.18B). This relatively small yet significant repression of the TWSG1

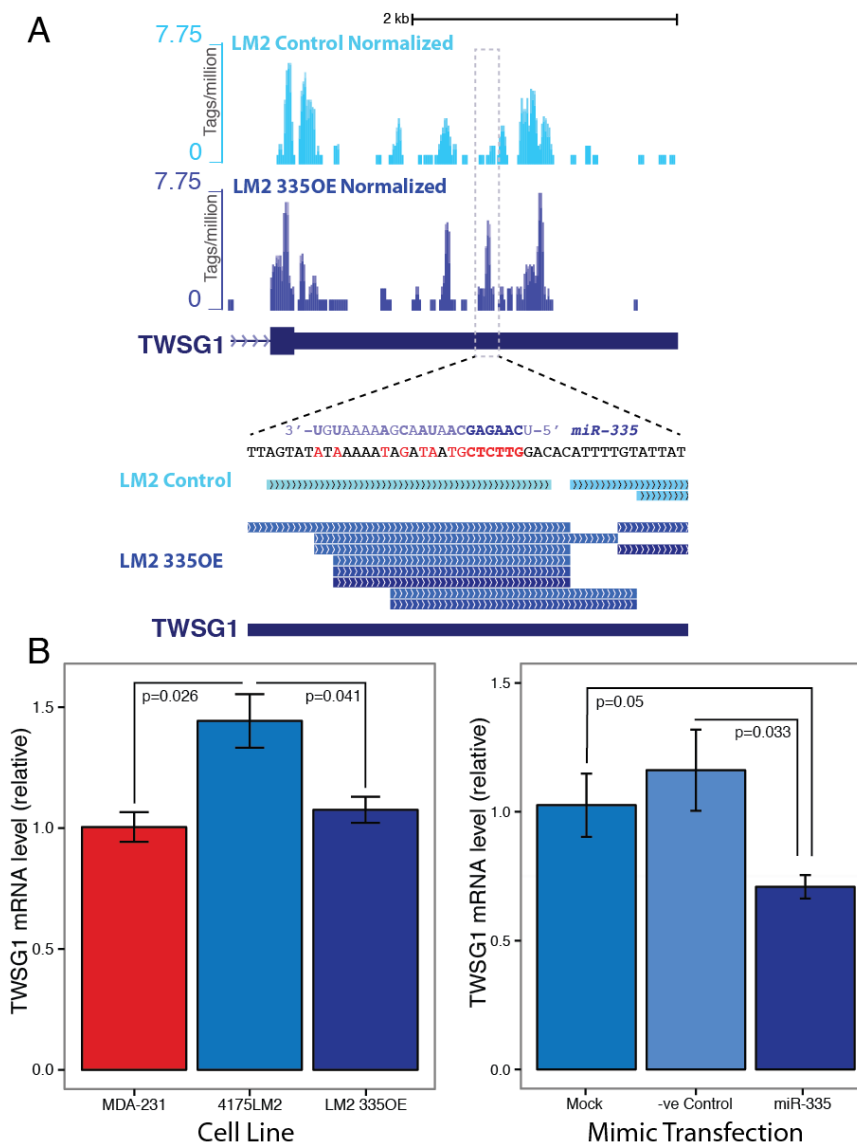


FIGURE 4.18: Differential AGO binding identifies TWSG1 as a miR-335 target. **A:** The *de novo* AGO binding site identified on the 3'UTR of TWSG1 contains including a miR-335 seed match in position 2-9. The miR-335 seed conservation is restricted to primates. A background of multiple common AGO binding sites is seen in other regions of the TWSG1 3'UTR. **B:** TWSG1 mRNA levels track with endogenous pre-miR-335 in MDA-231, 4175LM2 and are reduced to MDA-231 levels in LM2 335OE cells. TWSG1 mRNA is downregulated by transient transfection of miR-335 mimic into 4175LM2 cells. Mock is transfection reagent alone, negative control miRNA mimic is non-targeting in human. n=3 cell line, n=4 mimic transfection, expression levels normalized to TBP, error bars = s.e.m, p=values calculated by t-test.

mRNA failed to translate into reduced luciferase activity in TWSG1 3'UTR reporter assays with exogenous miR-335 mimic transfection (data not shown).

Relaxing the criteria for cluster selection identified a further two miR-335 and miR-335* targets regulated at the protein level. SDAD1 contains an AGO peak with log2 FC CLIP cluster binding of 2.84 (10 tags LM2 335OE, 2 tags LM2 Control, p-value=0.006) with complementarity to miR-335* in positions 2-7 with an additional G:U wobble at position 8. Luciferase reporter assays with WT full length SDAD1 3'UTR revealed exogenous mir-335* mimic repression of luciferase activity (Figure 4.19A) that was reproducible in replicate experiments. Similar results were seen with CDH11, a transcript with *de novo* AGO binding on a miR-335 seed (log2 FC CLIP cluster=3,32, p-value=0.010) and whose 3'UTR is regulated at the translational level by miR-335 mimic in luciferase reporter assays (Figure 4.18B). Both SDAD1 and CDH11 were unchanged at the mRNA level with miR-335/* mimic transfection when compared to mock control (data not shown) indicating that a portion of the *de novo* miR-335/* regulation identified by AGO HITS-CLIP may lead to translational repression of targets rather than transcript degradation.

Comparison of AGO HITS-CLIP derived de novo clusters miR-335/ targets and previously published targets*

Literature search identified multiple reported miR-335 targets in a variety of different cell types, including the LM2 335OE cells used in this

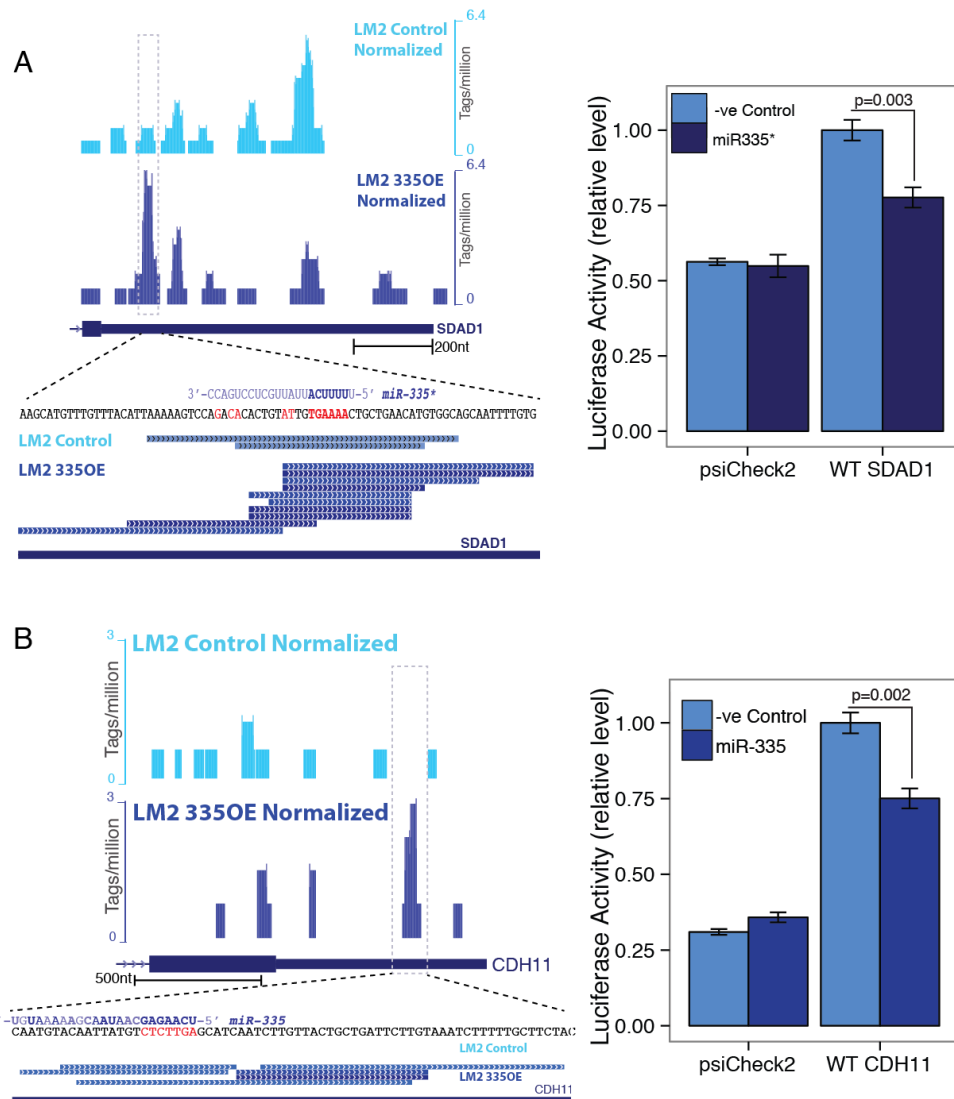


FIGURE 4.19: SDAD1 and CDH11 are regulated at the protein level by miR-335* and miR-335. **A:** Left - Genomic location of the pre-miR-335 dependent *de novo* AGO-mRNA binding site in the SDAD1 3'UTR, centered on a miR-335* seed site (2-7 match with G:U wobble in position 8). Right - Repression of luciferase reporter activity of WT full length SDAD1 3'UTR dual luciferase reporter construct when transfected into LM2 cells with 50nM miR-335* mimic, but not negative control mimic. **B:** As in A for miR-335 target CDH11. The reduction in luciferase activity for both 3'UTR constructs was observed in two independent experiments. All transfections n=4, mRNA expression levels normalized to TBP. Error bars = s.e.m, p-values calculated by t-test.

study (Table 4.6 with associated references). Of the previously published targets of miR-335, only DKK1 was identified by AGO HITS-CLIP as having gain of binding in LM2 335OE, although the miR-335/* seed sequence containing peaks identified here do not map to the non-conserved, non-canonical miR-335 seed site previously described (Zhang et al., 2011). Two miR-335/* dependent, LM2 335OE enriched AGO binding sites were identified in the 3'UTR of DKK1 by differential AGO HITS-CLIP (Figure 4.20A (raw tags) and B (normalized)). Significantly, the miR-335/* seed containing *de novo* binding events identified in LM2 335OE are also present in MDA-231 cells where miR-335/* is endogenously expressed.

Surprisingly, Q-PCR analysis of DKK1 steady state transcript levels revealed a strong decrease in 4175LM2 cells compared to MDA-231 parental cells that was not further decreased with the stable expression of pre-miR-335 in LM2 335OE cells (Figure 4.21A). Additionally, DKK1 mRNA levels were unchanged by the addition of exogenous miR-335/* mimics in 4175LM2 cells (Figures 4.21B& 4.21C). Similarly, full-length WT DKK1 3'UTR luciferase reporter constructs were unaffected by exogenous miR-335/* transfection in either MDA-231 or 4175LM2 cells (data not shown). It remains to be tested whether relieving miR-335/* binding of DKK1 by LNA inhibition of miR-335/* or target protection of the miR-335/*-dependent AGO-mRNA binding sites leads to derepression of the DKK1 transcript at either the mRNA or protein level in MDA-231 cells.

TABLE 4.6: Previously published miR-335 target sites compared to LM2 335OE AGO HITS-CLIP.

Cell Line	Gene	Evidence of AGO binding in LM2 335OE	Reference
MDA-231	SOX4	None	(Tavazoie et al., 2008)
	PTPRN2	None	
	MERTK	None	
U2OS	Rb	Weak, no differential binding	(Scarola et al., 2010)
MLO-A5	DKK1	miR-335 site not conserved in human Strong evidence of differential binding elsewhere on transcript	(Zhang et al., 2011)
hMSCs	RUNX2	None	(Tome et al., 2011)
Rat Kidney	SOD2	None, site not conserved	(Bai et al., 2011)
Malignant Astrocytoma	DAAM1	None	(Shu et al., 2011)
Human neuroblastoma	ROCK1	None	(Lynch et al., 2012)
	MAPK1	None	
	LRG1	None	
mESCs	OCT4	None	(Schoeftner et al., 2012)
	Rb	Weak, no differential binding	
Gastric Cancer	SP1	None	(Xu et al., 2012)
	Bcl-w	None	

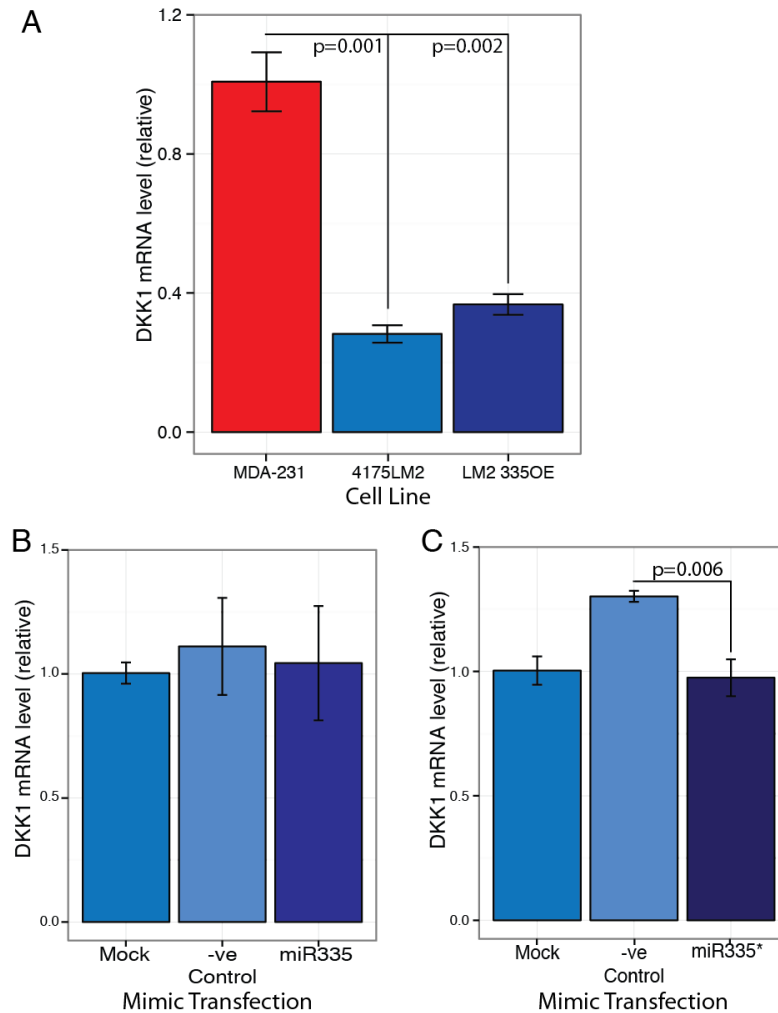


FIGURE 4.21: Endogenous DKK1 transcript levels are strongly repressed in LM2 cells compared to MDA-231 and show no further repression by exogenous miR-335/*. **A:** Q-PCR analysis revealed a significant decrease in the steady state levels of endogenous DKK1 in 4175LM2 cells compared to MDA-231 that was not further decreased in LM2 335OE. **B:** Transient transfection of miR-335 mimic had no effect on DKK1 levels in 4175LM2 cells. **C:** Transient transfection of miR-335* mimic had no effect on DKK1 transcript levels compared to mock transfected 4175LM2 cells. N=3 for each condition, mock transfections carried out with transfection reagent only, error bars = s.e.m, p-values calculated by t-test.

Multiple potential seed matches are present in the 62nt footprint of AGO on DKK1 in the region of significant peaks 1 and 2 (Figure 4.20B, right panel), however the majority of seed-matched miRNAs are either not present in the AGO-miRNA LM2 libraries, or are present at levels lower than miR-335 and miR-335*. In peak 2, no other miRNA with a seed match is present at higher levels than miR-335, while peak 1 contains seed matches to both miR-335* and the miR-374a/b family. miR-374a/b represents 2.40% of the LM2 Control miRNA profile and 3.24% of LM2 335OE, however this increase was not found to be significant by EdgeR analysis. Additionally, miR-374a/b is not significantly different between the MDA-231 and 4175LM2 AGO-miRNA profiles. This strongly suggests that the *de novo* binding identified by differential HITS-CLIP represents direct miR-335/* mediated AGO binding in the 3'UTR of DKK1, however the biological consequence of *de novo* miR-335/*-mediated AGO binding in the DKK1 transcript is currently unknown.

DISCUSSION

In an attempt to map on a genome-wide scale the targets of miR-335, a potent inhibitor of breast cancer metastasis, we have applied differential AGO HITS-CLIP to a previously characterized lung metastatic derivative of MDA-231 breast cancer cells, 4175LM2, over-expressing either a non-targeting control short hairpin or the pri-miR-335. Overexpression of miR-335 in 4175LM2 cells significantly reduces the invasive capability of these cells both *in vitro* and *in vivo* (Tavazoie et al., 2008). Comparative analysis of the AGO-miRNA HITS-CLIP profiles of LM2 335OE and LM2 Control indicates that pri-miR-335 overexpression does not globally alter the miRNA repertoire covalently bound to AGO in LM2 cells. Furthermore, despite over 100-fold increase in expression levels, we have found that miR-335 and miR-335* remain relatively minor components of the AGO-miRNA landscape.

In LM2 335OE, miR-335* represents 0.42% of the miRNA population and is the 21st most sequenced miRNA, with those miRNAs ranked above representing 92.9% of all AGO-miRNAs. This would suggest that despite using a strong promoter, pri-miR-335 is not expressed at supra-physiological levels, mitigating the effect of potential off-target effects through AGO saturation (Khan et al., 2009). Several rare miRNAs decrease in AGO-miRNA tag count in LM2 335OE indicating that some displacement of endogenous miRNA may occur with miR-335 overexpression. The relatively small shift in miR-335/* AGO-miRNA representation with stable overexpression of pri-miR-335 is in sharp

contrast to published data using transient transfection of miR-124 mimic into HeLa cells. In that system miR-124 becomes the 3rd most frequently cloned miRNA in transfected HeLa cells, a shift in the percentage of AGO HITS-CLIP miRNA reads from 0.15% to 8.4%, with the top two miRNAs comprising approximately ~32% of cloned AGO-miRNA (Chi et al., 2009).

The AGO-miRNA profile derived from LM2 335OE cells revealed an unexpected bias towards binding to the star, and presumed passenger strand of pre-miR-335. By convention mature miRNAs have been defined as the duplex strand that accumulates at a higher steady-state level than the partner strand, with bias observed towards strands with lower internal stability at the 5' end (Khvorova et al., 2003). Consistent with both these observations miR-335* contains a 5nt stretch of A-T base pairs at its 5' end compared to the more G-C rich 5' end of miR-335, and is consistently detected at ~2-3 fold higher levels in MDA-231 and all LM2 cells used in this study. Additionally, miR-335* is highly conserved, particularly in the region spanning the seed sequence, a feature which implies potential for regulatory function (Guo and Lu, 2010, Yang et al., 2011).

Microarrays used in the initial discovery of the anti-metastatic property of miR-335 did not contain probes for miR-335* as the sequence was not added to miRBase until a later date. It is clear from *in vivo* metastasis assays with LNA knockdown in MDA-231 cells, that miR-335 functions a mature targeting miRNA with anti-metastatic properties

separate from any potential concurrent miR-335* regulation of targets (Tavazoie et al., 2008). Both miR-335 and miR-335* have been shown to have equally repressive effects against perfect-match antisense luciferase reporter assays when simultaneously expressed from the pri-miRNA (Yang et al., 2011). Similar dual regulation of targets has been described for miR-199a-3p/-5p in the promotion of melanoma metastasis (Pencheva et al., 2012), miR-126/* in inhibition of breast cancer metastasis through modulation of the tumor microenvironment (Zhang et al., 2013), and miR-24-2/* in suppression of cell survival in MCF-7 breast cancer cells (Martin et al., 2012). Given the predominance of miR-335* *de novo* peaks identified by differential AGO HITS-CLIP cluster analysis in LM2 335OE cells, it is probable that miR-335* also regulates the metastatic ability of these cells.

13.4% of the pri-miR-335 dependent *de novo* AGO clusters identified contain a miR-335 or miR-335* 6nt seed sequence corresponding to position 1-8 located within the 62nt AGO footprint. This is lower than the proportion of *de novo* miR-124 canonical 6mer seed-containing sites detected in HeLa cells after mimic transfection (22.4% (Chi et al., 2009, Chi et al., 2012)), and miR-155-dependent AGO binding sites identified by differential AGO HITS-CLIP in WT and miR-155 knockout activated CD4+ T cells, (~57% (Loeb et al., 2012)). Positive correlation between the fraction of the AGO HITS-CLIP derived miRNA profile and the number of seed containing mRNA binding sites has been reported in multiple settings, including mouse brain, embryonic stem cells and CD4+ T cells, *C.elegans* and several cell lines (Chi et al., 2009, Hafner et al., 2010, Leung

et al., 2011, Loeb et al., 2012, Zisoulis et al., 2010). miR-155 and exogenous miR-124 mimic each represent a major component of the AGO-miRNA profile in their respective data sets, accounting for the high signal to noise in identification of directly targeted *de novo* binding sites.

The total number of miR-335/* targets identified by AGO HITS-CLIP in LM2 335OE, including the potential non-canonical bulged and G:U wobbles, is less than that reported for the most abundant miRNAs in previous AGO HITS-CLIP experiments (>300 miR-155 (Loeb et al., 2012), 691 miR-126 (Chi et al., 2009, Chi et al., 2012)). However, the number of miR-335/*-dependent seed containing AGO binding sites are similar in magnitude to the number of targets identified for less abundant miRNAs in mouse embryonic stem cells (Leung et al., 2011). The relatively minor combined contribution that AGO-miR-335/* make to the overall LM2 335OE miRNA profile is reflected in the low total tag numbers in AGO clusters delimiting functional binding sites. This represents an unanticipated limitation of the AGO HITS-CLIP technique in identifying statistically significant and robust miR-335/* mediated binding sites in this cellular context.

Despite low signal to noise, we have used differential AGO HITS-CLIP to successfully identify several previously unknown, direct targets of miR-335/* regulation. *De novo* cluster analysis identified CETN3 as a direct target of miR-335*, an interaction that was found to decrease both the mRNA levels of endogenous CETN3 and protein levels of reporter

constructs in miR-335* seed-dependent manner. Human CETN3 has described roles in centrosome duplication, nucleotide excision repair via putative interactions with xeroderma pigmentosum group C (XPC), and in the connecting cilium of photoreceptor cells (Cox et al., 2005, Dantas et al., 2011, Middendorp et al., 1997, Trojan et al., 2008). Human CETN3 is closely related to the *Saccharomyces cerevisiae* Cdc31 protein, which along with roles in cell cycle control and spindle body duplication (Paoletti et al., 2003, Schild et al., 1981), has been implicated in both protein degradation (Chen and Madura, 2008) and maintenance of genome integrity as part of the THP1-SAC3- SUS1-CDC31 complex coupling transcription and mRNA export (Gonzalez-Aguilera et al., 2008, Jani et al., 2009).

While the previously described cellular functions of human CETN3 do not mirror those of miR-335/* loss in metastasis (altered morphology and increased invasive capacity (Tavazoie et al., 2008)), other miR-335/* targets identified and verified in this study have been implicated in the regulation of breast morphology or enhanced invasiveness of tumor cells. For example, TWSG1, a secreted protein that functions as a BMP4 agonist in *Drosophila* and vertebrate embryos (Scott et al., 2001), is required for normal mammary gland ductal elongation and lumen formation (Forsman et al., 2013). Additionally, CDH11 (OB cadherin), a mesodermally expressed member of the cadherin superfamily (Hoffmann and Balling, 1995), is expressed in invasive breast cancer and positively regulates cell migration (Li et al., 2011c, Pishvaian et al., 1999, Sarrio et al., 2008).

We did not detect AGO-mRNA binding in the seed regions of any previously published miR-335 targets. SOX4, a previously validated miR-335 target in MDA-231 cells (Tavazoie et al., 2008), was not bound by AGO in any of the cell lines assayed by AGO HITS-CLIP in this study. The previously published miR-335 targeting of SOX4 was validated using dual luciferase reporter constructs encoding a 60nt region of the SOX4 3'UTR spanning a region of complementarity to miR-335 (TCTTGG, corresponding to a position 2-6 seed match with G-U wobble at position 1). Given the lack of biochemical evidence of AGO binding in the SOX4 3'UTR, the miR-335-dependent regulation of luciferase activity reported may reflect off-target effects rather than endogenous regulation in the context of the full length 3'UTR. Similar off target effects were seen with 3'UTR reporter assays using FAM105B, where both miR-335 and miR-335* mimics were able to repress luciferase activity due to the presence of seed matches in the 3'UTR, despite only miR-335* seed containing regions showing differential AGO binding in LM2 335OE. However, it is noted that lack of binding at the SOX4 3'UTR seen in AGO HITS-CLIP may also result from binding occurring below the threshold of detection of HITS-CLIP for a minor miRNA, and that the lack of binding at other previously validated miR-335 target sites may also reflect false negatives.

The apparent lack of miR-335 or miR-335* mediated regulation of DKK1 is difficult to reconcile with the robust, pre-miR-335 dependent binding seen on the DKK1 3'UTR. No other miRNAs with canonical seed

matches in the 62nt AGO footprint are expressed at higher levels than miR-335/* in either MDA-231, LM2 Control or LM2 335OE cells nor do they show significant differences in AGO-miRNA counts that correlate with the gain and loss of AGO-mRNA binding observed between cell lines. The regulation by miR-335/* may be below the threshold of detection by the assays used in this study, or the 3'UTR of DKK1 may potentially function as a neutral target with respect to miR-335/* regulation. AGO binding to the DKK1 3'UTR may sequestering miR-335/* in MDA-231 and LM2 335OE allowing it to function as a competitive endogenous RNA by (Salmena et al., 2011, Seitz, 2009). Interestingly, competitive, DICER-dependent antagonistic regulation has been documented between the DKK1 and PTEN 3'UTRs in diabetic cardiomyocytes through sequestration of miR-93 and miR-106a (Ling et al., 2013).

The ability of endogenous non-coding transcripts and coding mRNA 3'UTRs to influence tumorigenesis, presumably through titration of miRNAs or other regulatory RNA binding proteins, is being increasingly recognized. miRNA-mediated regulation of the tumor suppressor PTEN can be modulated by the expression of the pseudogene PTENP1 (Poliseno et al., 2010) or concomitant expression protein-coding transcripts targeted by common miRNAs in a DICER dependent manner (Tay et al., 2011). Similarly the 3'UTRs of versican (Fang et al., 2013a), prohibitin (Manjeshwar et al., 2003), α -tropomyosin (Rastinejad et al., 1993) and CD44 (Rutnam and Yang, 2012) have been shown to regulate

tumorigenesis and metastasis in various cancer types, separating the regulatory function of non-coding regions of transcripts from any protein-mediated effect.

Loeb et al. (2012) have suggested that the relatively few highly bound targets of miR-155 identified by AGO HITS-CLIP would argue against the endogenous miRNA sponge hypothesis, as very few transcripts appear to bind the miRNA-AGO complex at high enough levels to lead to sequestration. However, given that the threshold of regulation by miRNAs has been shown to be determined by both the level of target and miRNA (Mukherji et al., 2011), regulation by a low abundance miRNA such as miR-335/* may be more sensitive to endogenous sponges. The decrease in the steady state levels of DKK1 mRNA in lung metastatic derivatives occurs concomitantly with copy number loss in the miR-335 locus, increased MEST promoter methylation in LM2 cells and increased invasive capacity (Png et al., 2011). If DKK1 3'UTR functions as a sponge for miR-335/*, overexpression of this region of the DKK1 transcript should enhance the invasive capacity of both MDA-231 and LM2 335OE through sequestration of miR-335/* away from its endogenous targets, while mutant constructs with disrupted seed sites should have no effect on invasion.

An increasing body of work implicates altered miR-335 expression in multiple tumor types, with low miR-335 levels detected in breast, ovarian, neuroblastoma, and gastric cancer (Png et al., 2011, Polytaichou

et al., 2012, Sorrentino et al., 2008, Tavazoie et al., 2008, Lynch et al., 2012, Xu et al., 2012), and miR-335 up regulation reported in malignant astrocytoma and meningioma (Shi et al., 2012, Shu et al., 2011). While experimental validation of the role of miR-335* in *in vivo* metastasis is currently on going, our discovery that miR-335* can function as a guide strand opens up new directions in the search for regulated downstream transcripts and for potential for therapeutic intervention in multiple cancer settings. The data we have presented provides a first look at the network of transcripts regulated by both miR-335 and miR-335* in breast cancer, and provides direct biochemical evidence of multiple new targets with as yet unknown functions in breast cancer metastasis.

CHAPTER 5: CHARACTERIZATION OF RBM47, AN RNA-BINDING SUPPRESSOR OF BREAST CANCER METASTASIS

INTRODUCTION

High-throughput crosslinked immunoprecipitation (HITS-CLIP) was initially developed to identify *in vivo* directly bound targets of the neuron-specific alternative splicing regulator Nova (Ule et al., 2003, Ule et al., 2005b, Licatalosi et al., 2008). Since UV-irradiation leads to covalent bond formation only between proteins and RNA that are in direct contact (within Angstroms (Darnell, 2010)), the resulting sequence data can be used to determine both the target RNAs bound by an RNA binding protein (RBP) and the specific region within the transcript that is bound. Furthermore, HITS-CLIP data can be used to validate the sequence specificity of the protein in question with single-nucleotide resolution (Hafner et al., 2010, Zhang and Darnell, 2011). The HITS-CLIP technique has now been applied to multiple RBPs, particularly those involved in neurological diseases, to identify both bona fide disease-related sets of RNA targets and to elucidate normal biological functions.

The genome-wide nature of the data derived from HITS-CLIP has lead to discovery of new regulatory functions for previously well-characterized proteins indicating that the technique can also be use prospectively. For example, Nova HITS-CLIP has greatly expanded the repertoire of Nova-regulated alternative splicing events to >600 (Zhang et al., 2010a), and lead to the identification of a previously unknown role for

Nova in the regulation of alternative polyadenylation (Licatalosi et al., 2008) and steady-state target expression levels through direct regulation of NMD exons (Eom et al., 2013). Additionally, CLIP has revealed unexpected roles in miRNA maturation for hnRNP-A1. This extensively studied RNA splicing regulatory protein was found to directly bind the stem-loop sequences in pri-miR-18a to facilitate Drosha-mediated processing (Guil and Caceres, 2007, Michlewski et al., 2010)

In this study, we propose using HITS-CLIP to identify the targets and elucidate the biological function of a previously uncharacterized, presumed RNA binding protein with direct links to breast cancer metastasis. RNA binding motif protein 47 (RBM47) is so named due to the predicted presence of three RNA recognition motifs (RRM domains). The combination of one or more RRM domains allows for recognition of single stranded RNA sequences of 8-10 nucleotides, although RRM domains also been demonstrated to mediate protein-protein interactions (Maris et al., 2005). The previously mentioned ELAVL proteins all contain three RRM domains. Increased expression of the ubiquitously expressed ELAVL family member ELAVL1 (HuR) has been shown to promote a more tumorigenic phenotype in xenograft models of breast cancer (Mazan-Mamczarz et al., 2008a), and has been associated with poor prognosis in multiple cancers, including invasive ductal breast carcinoma (Denkert et al., 2004, Heinonen et al., 2005).

RBM47 was first identified by Ding et al, 2010 as a gene of interest in the metastatic progression of breast cancer. Deep sequencing of genomic DNA derived from a patient's primary basal-like aggressive tumor and subsequent cerebellar metastases revealed a point mutation in the coding sequence of RBM47 that was highly enriched in brain metastases (79.15% mutant allele frequency, FDR=0.03). This point mutation leads to a G-deletion at nucleotide position 1280, a frame shift and the introduction of a premature stop codon (Ding et al., 2010) suggesting that loss of wild type function of RBM47 (or gain of function in the mutant allele) was selected for during metastasis to the brain. Additionally, RBM47 expression levels have been shown to decrease over ten-fold during the epithelial-mesenchymal transition (EMT) of human mammary epithelial cells expressing a tamoxifen-inducible Twist-ER fusion protein (Shapiro et al., 2011).

RBM47 is currently an uncharacterized protein; therefore we propose using HITS-CLIP to determine its RNA binding abilities. Furthermore, by mapping robust RBM47 interactions on target RNAs we aim to identify both the functional of RBM47 in RNA metabolism and to identify targets through which RBM47 mediates its proposed anti-metastatic functions.

RESULTS

RBM47 expression is decreased in the metastatic derivatives of MDA-231

Re-analysis of published microarray data from MDA-231 and its metastatic derivatives with tissue-specific metastasis to lung (4175LM2), bone (1833BoM2) and brain (831BrM2) revealed down regulation of RBM47 in all metastatic sub lines (data not shown) (Bos et al., 2009, Kang et al., 2003, Minn et al., 2005a). In particular, RBM47 expression is dramatically decreased in 831BrM2 (Figure 5.1A), with corresponding decrease in protein expression as assayed by Western blot (Figure 5.1B). RBM47 protein levels were also lower in lung and bone metastatic derivatives when compared to the parental MDA-231 cell lines suggesting that downregulation of RBM47 is a common feature associated with increased metastatic potential. Data analysis and Western carried out by Sakari Vanharanta, Joan Massagué laboratory, MSKCC.

Wild-type RBM47 functions as a metastasis suppressor

Doxycycline-inducible 831BrM2 triple reporter stable cell lines were generated that express either wild type FLAG-tagged RBM47 (Tet-On Flag-RBM47 WT) or a mutant RBM47 harboring the same frame shift mutation identified by Ding et al., 2010 (Tet-On Flag-RBM47 MUT). Mice were injected in the left ventricle with of 50,000 Tet-On Flag-RBM47 WT or MUT cells, and fed a diet containing doxycycline (dox) starting three days after injection. Metastatic spread to the brain was assayed by *ex vivo* photon flux six weeks post injection. In control animals that were injected with Tet-On Flag-RBM47 WT cells but did not receive dox in food, a high

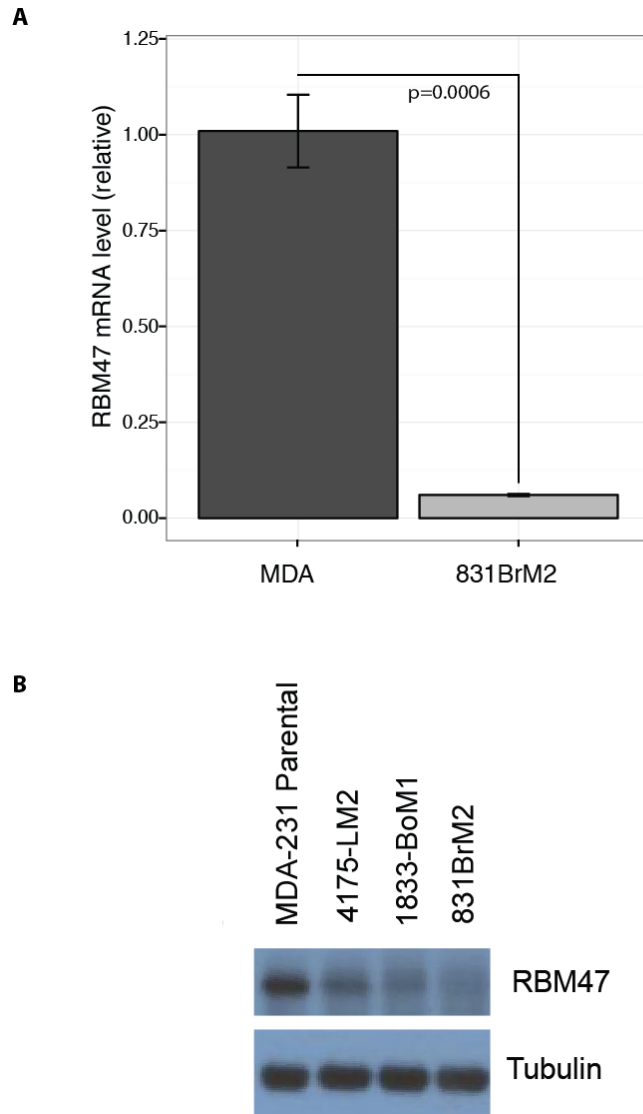


FIGURE 5.1: RBM47 expression is decreased in the highly metastatic derivatives the breast cancer cell line MDA-MB-231. A: Q-PCR showing decreased RBM47 transcript levels in 831BrM2 cells compared to MDA-231. Expression level normalized to TATA box binding protein (TBP). $n=3$, error bars = s.e.m, p-values calculated by t-test. **B:** Western-blotting reveals decreased RBM47 protein levels in the highly metastatic derivatives of MDA-MB-231 with tissue specificity towards the lung (4175LM2), bone (1833BoM1) and brain (831BrM2). Loading control tubulin. Western blot by Sakari Vanharanta, Joan Massagué laboratory, MSKCC.

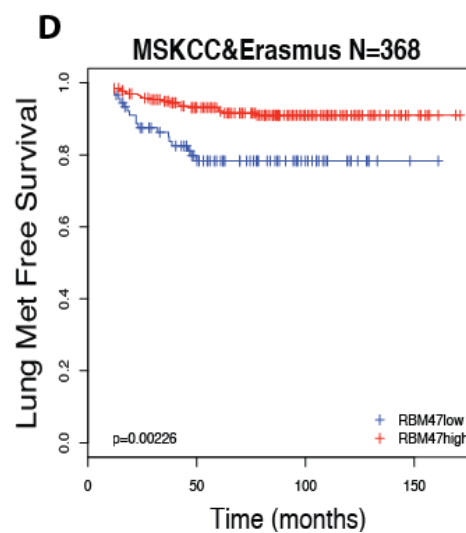
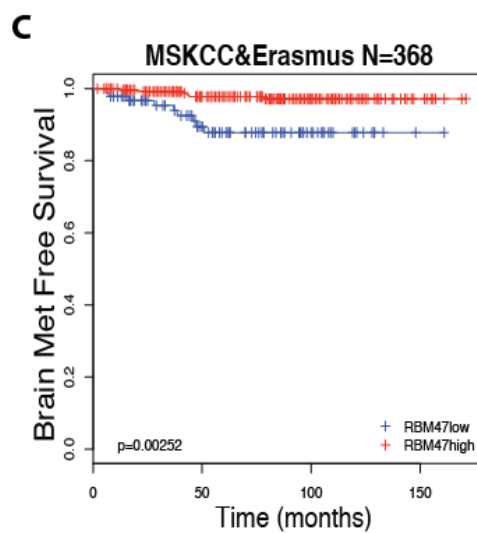
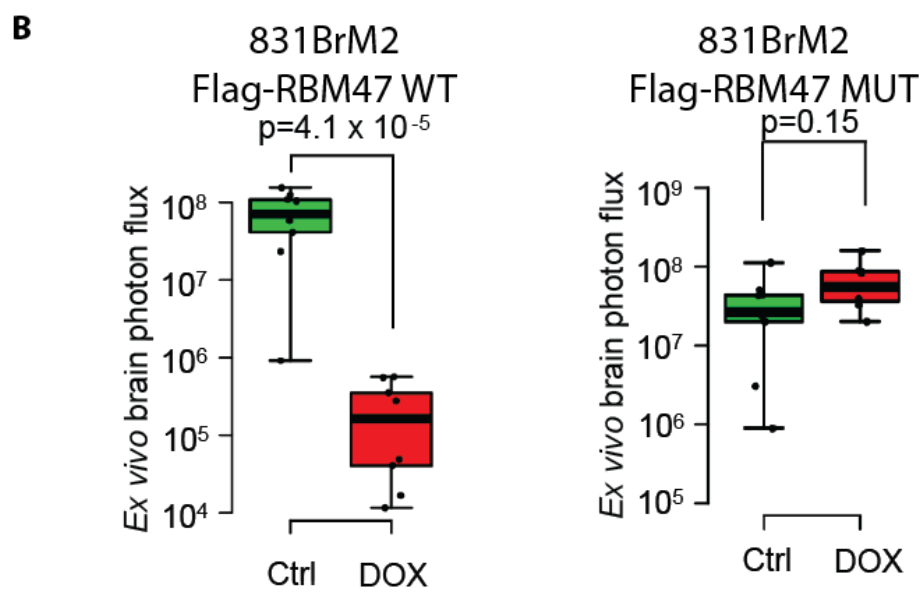
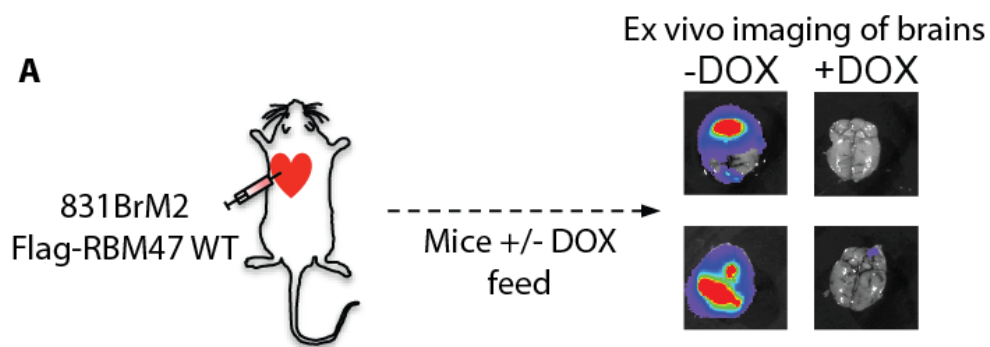
degree of brain metastasis was evident that was suppressed by dox-induced expression of WT RBM47 (Figure 5.2A and 5.2B). Expression of Mutant RBM47 had no effect on the metastatic potential of 831BrM2 cells (Figure 5.2B). RBM47 expression levels were found also to be clinically associated with human metastasis, with patients whose primary tumors expressed low levels of RBM47 having significantly shorter time to distal metastatic relapse to both brain ($p=0.0025$) and lung ($p=0.0023$, Figure 5.2C). All experiments and data analysis carried out by Sakari Vanharanta, Joan Massagué laboratory, MSKCC.

Domain architecture of RBM47

The RBM47 gene is located on the negative strand of chromosome 4 at position 40,120,029-40,326,640. Refseq indicates two alternatively spliced transcripts that are translated into 593 and 524 amino acid proteins respectively (Figure 5.3A). Conserved domain analysis shows the presence of three RNA recognition motif (RRM) domains in both RBM47 isoforms, with a variable region occurring between amino acids 374 and 414, C-terminal to the RNA binding motifs that corresponds to the alternatively spliced exon. RRM1 and RRM2 contain RNP-1 consensus motifs (RGYSFVMY and RGFAFVEY respectively) while RRM3 contains an RNP-2 consensus motif (LYVRNL). The mutant RBM47 identified in patient cerebellar metastases produces a truncated protein containing only RRM1 and 33 amino acids of RRM2 (Figure 5.3B).

FIGURE 5.2: RBM47 functions as a metastasis suppressor *in vivo*, with low expression clinically associated with poor metastasis-free survival.

A: Mice were injected with clonal populations of BrM2 cells stably expressing either WT or mutant RBM47 then fed a diet plus/minus doxycycline to induce transgene expression. At 6 weeks post injection, brains were imaged *ex vivo* to assess level of metastasis (photon flux). **B:** Dox-induced expression of WT RBM47 in 831BrM2 inhibits metastatic colonization of the brain *in vivo* when compared to non-dox treated animals (6 weeks post intercardiac injection). Expression of Mutant RBM47 has no effect on the metastasis (n=8). Kaplan-Meier curves for the combined Memorial Sloan Kettering and Erasmus Medical Center breast tumor cohorts (368 tumors) showing brain (**C**) and lung (**D**) metastasis free survival of patients whose primary tumors expressed low levels of RBM47. All experiments carried out by Sakari Vanharanta, Joan Massagué laboratory, MSKCC.



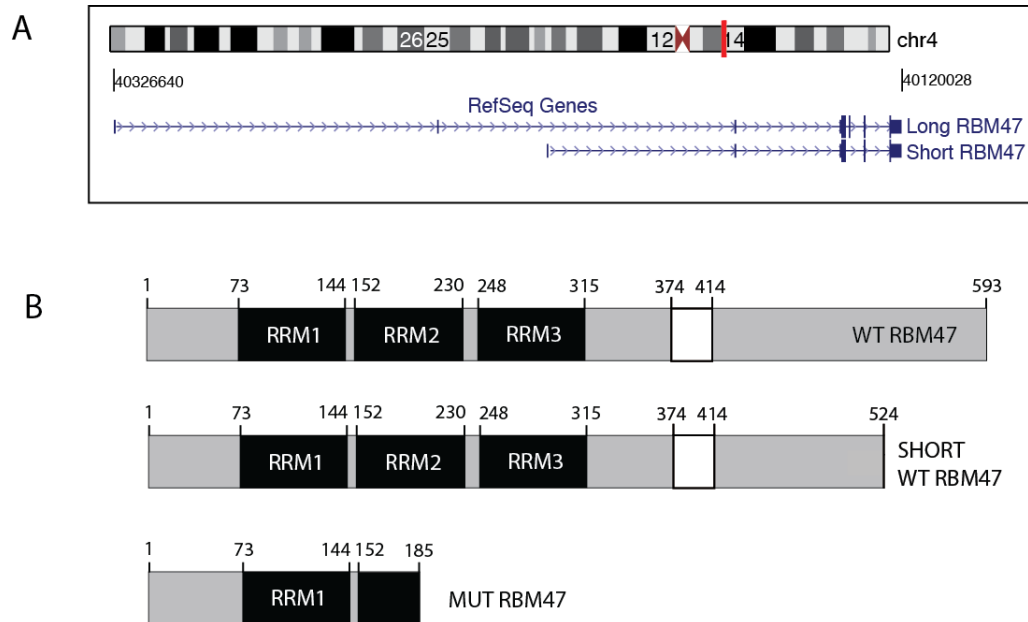


FIGURE 5.3: RBM47 gene and protein domain organization. **A:** The RBM47 gene encodes two RefSeq annotated isoforms with alternative 5'UTRs and one alternative coding exon. The longer isoform of RBM47 was used for all subsequent analyses of RBM47 function. **B:** Conserved domain analysis reveals the presence of three RRM domains (black) in both long and short isoforms of RBM47, with a variable region (white) corresponding to the alternative exon. Mutant RBM47 produces a truncated protein due to a G deletion at position 1280 and subsequent frame shift that introduces a premature stop codon in RRM2.

Protein BLAST of RBM47 reveals sequence similarities to the known RNA-binding proteins

Protein BLAST of RBM47 reveals high sequence identity and similarity with both apobec-1 complementation factor (A1CF, 58.5% and 72.65%, respectively) and hnRNP-Q/SYNCRIP (55.3% and 73%), proteins that also contain three RRM domains. Identity and similarity were calculated using the EBOSS6.3.1 matcher (Rice et al., 2000). RBM47 alignments with A1CF and hnRNP-Q (generated by ClustalW2 (Goujon et al., 2010)) are shown in Figures 5.4. High sequence identity is evident in the N-terminal regions spanning the RRM domains (red dashed line), with more sequence divergence at the C-terminal ends. Both A1CF and hnRNP-Q have been shown to directly bind RNA and to regulate RNA editing (Mehta et al., 2000, Blanc et al., 2001). Additionally, hnRNP-Q has multiple functions in RNA metabolism including RNA localization, splicing and stability (Bannai et al., 2004) (Chen et al., 2008) (Weidensdorfer et al., 2009). Together this would strongly suggest that the RRM domains in RBM47 are capable of binding RNA.

RBM47 is an RNA binding protein

RBM47 has been identified as a component of the protein-mRNA interactome in two independent studies of oligo(dT) affinity purified mRNA-RBP complexes in HeLa and HEK293 cells (Baltz et al., 2012, Castello et al., 2012) although direct RNA binding activity has not been confirmed. To test whether RBM47 can directly bind RNA, we made use of the ability of UV-irradiation at 254nm to induce chemical crosslinks

between RNA and proteins that are in direct contact (on the order of Angstroms apart (Darnell, 2010)). Commercially available RBM47 antibodies were unsuitable for immunoprecipitation due to cross-reaction with multiple proteins, therefore 831BrM2 Tet-On empty vector control and 831BrM2 Tet-On Flag-RBM47 WT stable cell lines (generated by Sakari Vanharanta in the Massagué laboratory and henceforth referred to as Tet-On and RBM47-WT) were grown for 3 days in the presence of 1ng/ml doxycycline before UV-irradiation and harvesting. Non-crosslinked RBM47-WT cells grown in the same manner were used as a negative control. Lysed cells were RNaseA digested to reduce the modal size of any bound RNA to that of the footprint protected by protein binding, and stringently immunoprecipitated using an anti-Flag antibody. Bound RNA was end-labeled with γ -³²P-ATP using T4 PNK, protein separated by SDS-PAGE and complexes transferred to nitrocellulose.

Exposing the nitrocellulose to x-ray film (Figure 5.5A) revealed radioactive signal at the predicted size of RBM47 only in lanes corresponding to dox-treated, UV-irradiated RBM47-WT cells. The lanes for non-irradiated RBM47-WT and the UV-treated Tet-On Control cells were clear of radioactivity indicating UV-dependent RNA immunoprecipitation and lack of non-specific antibody pull down. Western blotting of the nitrocellulose membrane with a second Flag antibody (Figure 5.5B) showed no signal in the Tet-On control lanes indicating the specificity of the Flag antibody, and strong, equal intensity protein signal at the predicted size of RBM47 in both lanes corresponding

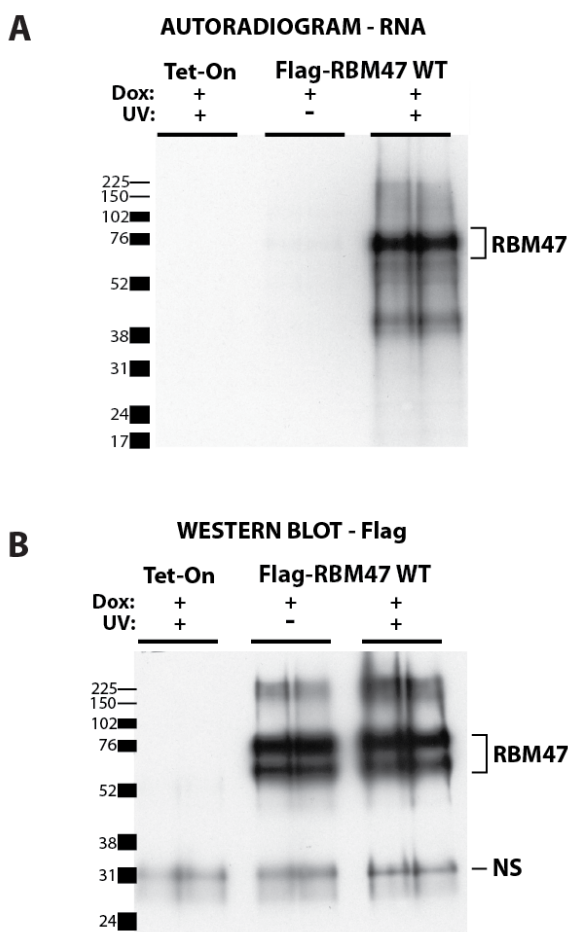


FIGURE 5.5: RBM47 is an RNA binding protein. **A:** 831BrM2 cells stably expressing tet-inducible WT Flag-RBM47 (RBM47-WT) or empty vector (Tet-On) were UV-irradiated before lysis and stringent anti-FLAG IP. Bound RNA was end-labeled with γ - ^{32}P -ATP and IPed protein size separated via SDS-PAGE. Transfer to nitrocellulose and subsequent exposure to x-ray film revealed the presence of radioactively labeled RNA only in those lanes corresponding to WT-RBM47 crosslinked cells, indicating that RNA pull down was both UV and FLAG epitope dependent. **B:** Western blotting of the same nitrocellulose membrane reveals the presence of FLAG-RBM47 protein in both cross-linked and non-crosslinked lanes at the same molecular weight as the radioactive RNA.

to dox-treated RBM47-WT. A second flag-reactive band is seen at a lower molecular weight and is presumed to be a degradation or cleavage product of the full length RBM47-WT. This experiment indicates the presence of RNA in physiologically relevant contact with RBM47 *in vivo*, and confirms RBM47 as an RNA binding protein.

Mapping direct RBM47-RNA interactions using a modified HITS-CLIP protocol

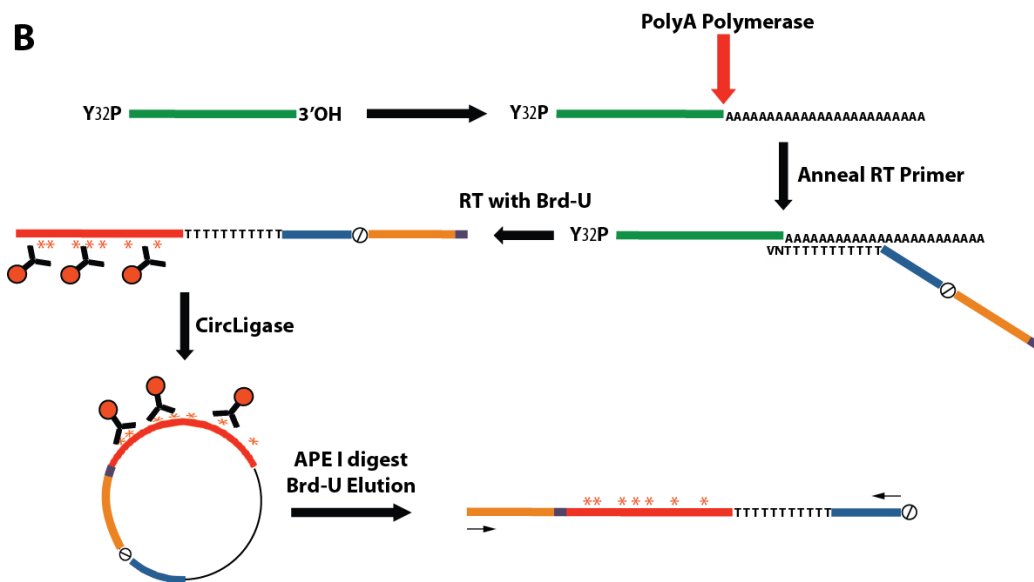
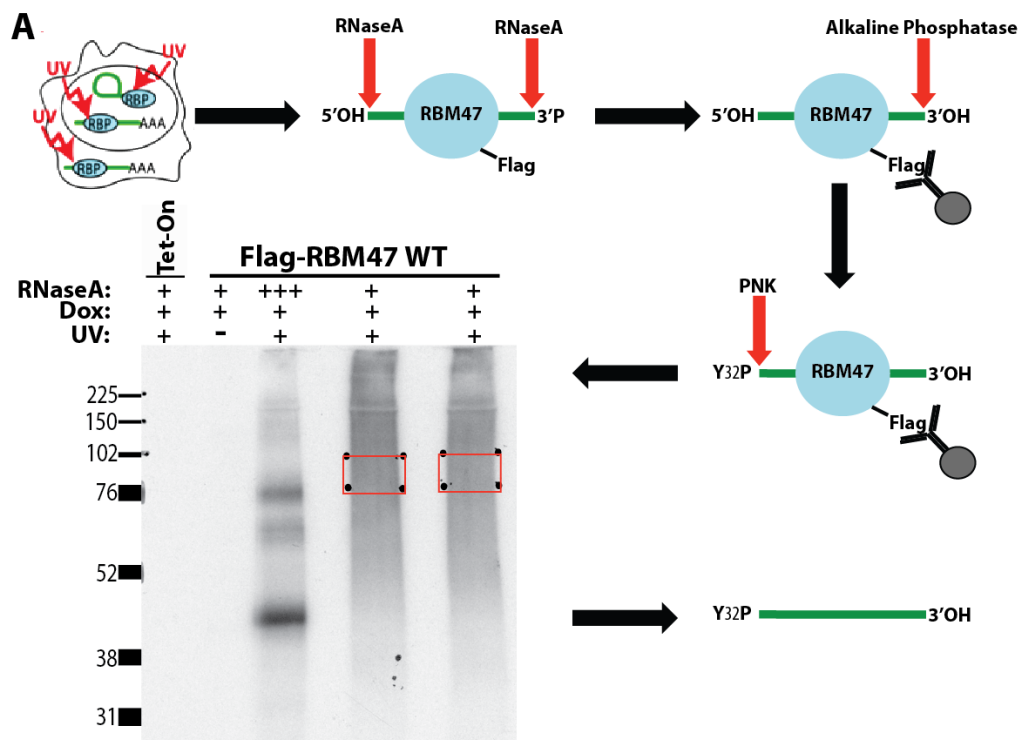
To identify direct RBM47 targets in 831BrM2 cells, a modified version of the HITS-CLIP protocol was employed that was developed by Aldo Mele in the Darnell laboratory (Figure 5.6). This protocol combines polyA tailing and a specially designed RT-PCR primer that reduces the number of linker ligation reactions while also utilizing cDNA circularization to capture fragments that may prematurely truncate during RT at the crosslink site (Konig et al., 2010). Both of these enhancements can theoretically increase the complexity of the sequenced HITS-CLIP libraries by reducing linker-ligation bias (Jayaprakash et al., 2011), removing inefficient ligation steps, and by capturing fragments that would normally not be PCR amplified due to lack of 5' complementary sequences.

Duplicate dishes of RBM47-WT and Tet-On Control cells were grown in the presence of 1ng/ml dox for 3 days before UV-crosslinking and harvesting for HITS-CLIP. A triplicate dish was harvested without crosslinking for RNAseq analysis. The Flag-RBM47 HITS-CLIP protocol differs slightly from the previously used AGO HITS-CLIP protocol in that

anti-Flag antibody was used for IP, and 3' linker ligation was omitted in favor of CIP treatment and 5' end labeling with γ -³²P-ATP (Figure 5.6A). As seen in Figure 5.5A, the autoradiogram shows no radioactive signal in both the Tet-On control and the non-crosslinked RBM47-WT control lanes while high levels of RNaseA treatment in crosslinked RBM47-WT cells reveals a sharp radioactive RNA band at the predicted size of RBM47. Reducing the concentration of RNaseA added pre-IP to crosslinked RBM47-WT lysates lead to a smear of radioactivity rising above the predicted weight of RBM47, corresponding to RNA-protein complexes of increasing RNA modal size. The red boxes indicate the region of the nitrocellulose that was excised (~30kD, equivalent to ~100nt RNA), proteinaseK treated, and from which the two RBM47-bound CLIP libraries were independently prepared.

In order to clone the RBM47-bound RNA library and prepare it for Illumina high throughput sequencing (Figure 5.6B), the isolated RNA fragments (green bar) were first polyA tailed then reverse transcribed using a primer with several important features. At the 5' end of the RT primer are two hexamers (purple bar); the first is a known-sequence 6-mer index used for identification of multiplexed samples, the second is a degenerate 6-mer barcode used to distinguish unique cDNA clones from PCR duplicates. Downstream from the index and barcode is a 21nt partial reverse complement to the Illumina forward sequencing primer (orange bar), an abasic furan that serves as an ApeI cut site (x), and the full sense sequence of the Illumina reverse sequencing primer (blue bar).

FIGURE 5.6: Outline of the modified HITS-CLIP protocol used with RBM47. **A:** 831BrM2 Flag-RBM47 WT and Tet-On control cells were treated with 1ng/ml doxycycline for 3 days before UV irradiation to induce covalent bonds between direct RNA-protein interactions. After cell lysis and partial RNaseA digestion, RBM47-protein complexes were stringently purified by IP with an anti-Flag antibody followed by CIP treatment and radioactive labeling with 5' $\gamma^{32}\text{P}$ -ATP. Complexes were separated by SDS-PAGE and transferred to nitrocellulose. Exposure of the nitrocellulose membrane to x-ray film revealed the presence of radioactive RNA in only those lanes with both Flag-RBM47 expression and UV-crosslinking. Treatment with high RNaseA (+++) reduced the radioactive signal to a sharp band at the anticipated molecular weight of Flag-RBM47. Low RNaseA treatment (+) of WT-RBM47 crosslinked cells leads to a smear of radiation above RBM47 indicating the presence of RBM47-RNA complexes of increasing RNA length. Small sections of nitrocellulose were excised from two independent replicate experiments as indicated by red boxes, and RBM47 bound RNA purified from the membrane by proteinase K digestion. **B:** Purified RBM47 CLIP tags were polyA tailed then reverse transcribed in the presence of BrdU, using a polydT-NV primer containing Solexa primer sequences (forward – orange, reverse – blue) separated by an abasic furan (\odot), with a short 6nt degenerate region plus a 6nt index sequence (purple). The resulting BrdU-cDNA was stringently purified by IP, circularized and digested with *ApeI* to re-linearize the cDNA fragment with Solexa sequences now flanking the original cloned RBM47-bound fragment. These cDNAs were then PCR amplified, multiplexed and high-throughput sequenced.



This is then followed by a stretch of T₂₀ plus a 3' VN to allow for anchoring of the RT primer at the 5' end of the polyA tail partially overlapping into the purified RNA fragment. RT-PCR was carried out in the presence of Brd-U (red *) to allow for stringent purification of the resulting cDNA (red bar) from the RT reaction by anti-Brd-U IP.

While still on-bead, CircLigase was used to catalyze intramolecular ligation, bringing the sequences needed for Illumina sequencing into the correct orientation with respect to the cloned fragment, so that upon ApeI linearization and Brd-U elution cDNA fragments are ready for PCR amplification and deep sequencing. Unpublished data from our laboratory has shown that more complex CLIP libraries can be generated from samples with low total PCR amplification, therefore having a single PCR step before Illumina sequencing in this modified protocol further enhances the complexity of the final data set.

Summary of RBM47 HITS-CLIP data

The processing and bioinformatic manipulation of the data obtained from RBM47-WT HITS-CLIP are summarized in Figure 5.7A. Briefly, sequencing of the two pooled RBM47-WT HITS-CLIP libraries gave a total of ~38.9 million reads. Reads were filtered based on quality score, exact duplicates collapsed, and adapter sequences and 3' polyA stretches trimmed. Alignment to the human genome (hg18) left ~14.0million reads mapping to single locations. Splitting the data by index assigned ~4.4million reads to replicate A and ~9.6million reads to

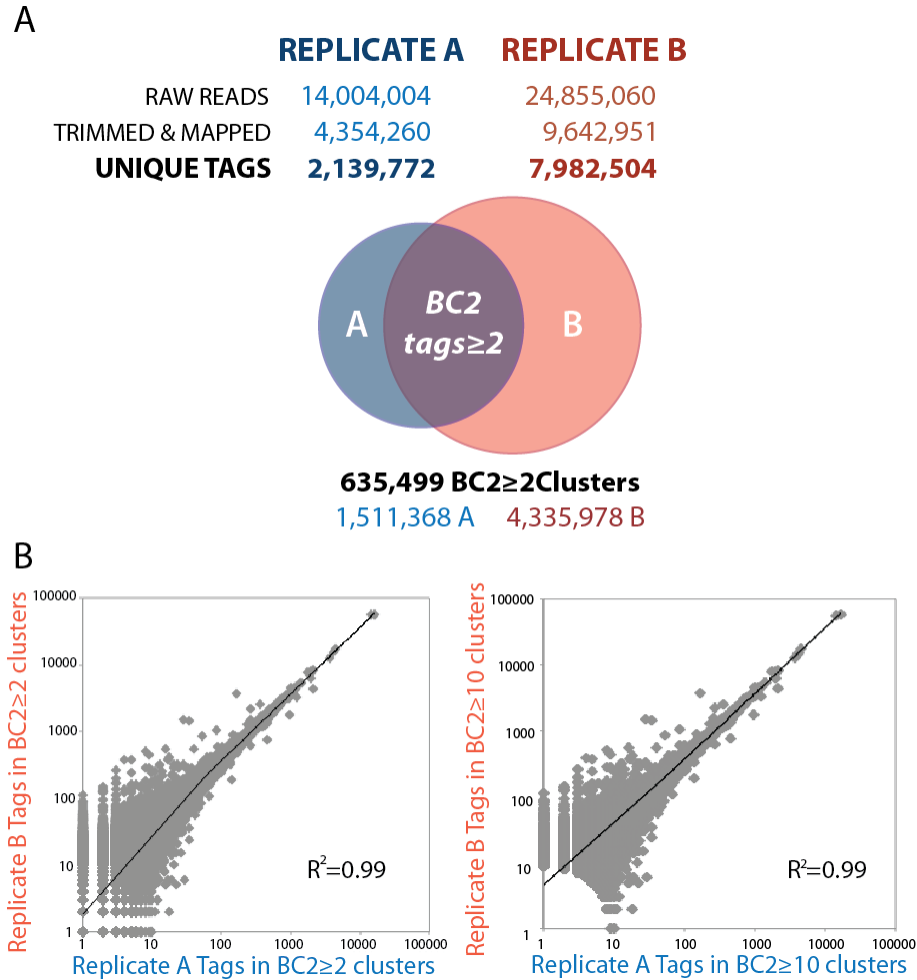


FIGURE 5.7: Summary of RBM47 HITS-CLIP sequencing data and reproducibility. **A:** Multiplexed sequences were filtered based on quality score and exact sequence duplicates collapsed. Adapter sequences were trimmed from both 5' and 3' ends, and A nucleotides trimmed from the 3' end. Sequences originating from each replicate library were separated using the 6nt index and reads aligned to hg18. Collapsing sequences based on degenerate linker and genomic position generated ~1.5million unique RBM47 tags for replicate A, and ~4.3 million for replicate B. **B:** Unique tags from each replicate were pooled together to define regions of reproducible RBM47 binding (BC2 \geq 2 tags). Plotting the number of tags each replicate contributes to a cluster reveals an extremely high degree of correlation, indicating a high degree of reproducibility between data sets. This correlation remains when increasing the stringency of cluster definition to BC2 \geq 10 tags.

replicate B. Sequences were further collapsed based on degenerate linker sequence and genomic start site to remove PCR duplicates, leaving ~2.14million Replicate A and ~7.98million Replicate B unique reads, referred to as RBM47 tags.

Reproducibility of RBM47 HITS-CLIP data sets

To identify reproducible sites of RBM47-RNA binding, all RBM47 tags were clustered together to return binding sites with greater than or equal to 3nt overlap between tags, and the minimal presence of one tag from each replicate library (biological complexity=2 and tags \geq 2 or BC2 \geq 2tags, minimal definition of a cluster). 635,499 BC2 \geq 2tags clusters were identified on 10,547 genes (55% of annotated genes). Replicate experiments were found be highly reproducible with strong correlation between the number of tags per BC2 \geq 2tags cluster from each experiment. This strong correlation between replicate libraries remained constant when cluster definition stringency was increased to BC2 \geq 10tags (both $R^2=0.99$, Figure 5.7B).

Genomic distribution of RBM47 binding

While requiring reproducibility between replicates at any given binding site reduces the potential for background noise in CLIP data sets, multiple additional measures are routinely applied to increase stringency of cluster definition. These include increasing the number of overlapping tags (BC2 \geq 10tags, 98,298 RBM47 clusters) or applying the previously described significant peak threshold (statistically significant observed

binding above that expected by random distribution of CLIP tags on a transcript; BC2 sigpeak \geq 10tags, 30,277 peaks). Another measure of robustness can be calculated by chi-squared analysis of tags per cluster per experiment (detailed in (Darnell et al., 2011)) then applying an empirical threshold on the returned p-value (BC2 χ^2 p \leq 0.01). For RBM47 HITS-CLIP the minimal tag number per cluster that passes this threshold was 26, leaving 19,908 clusters. All resulting clusters were annotated and the percentage of clusters per genomic location represented in a pie chart (Figure 5.8A). The same increasing stringency filtering criteria were applied to the previously described 831BrM2 AGO-mRNA HITS-CLIP clusters for comparison (Figure 5.8B).

Annotation of BC2 \geq 2tags clusters revealed predominant RBM47 binding in intronic regions (53%) followed by deep intergenic sequences (21%). Binding to deep intergenic regions is considered to be a combination of both real, unannotated transcribed targets and noise, since many of the clusters result from mapping errors to low complexity sequences and repeat elements. Given the uncertain nature of deep intergenic binding events they were omitted from further analysis. The remainder of BC2 \geq 2tags clusters segregate almost equally into 3'UTR (9%), the 10kb sequence immediately downstream of annotated 3'UTR ends (presumed to be unannotated alternative 3'UTRs, 8%), and coding sequence (7%). Only 2% of BC2 \geq 2tags clusters map to 5'UTR regions (Figure 8A). A similar distribution of binding is seen for AGO in 831BrM2 cells at this low stringency cluster threshold, BC \geq 2 tag \geq 2. Increasing the

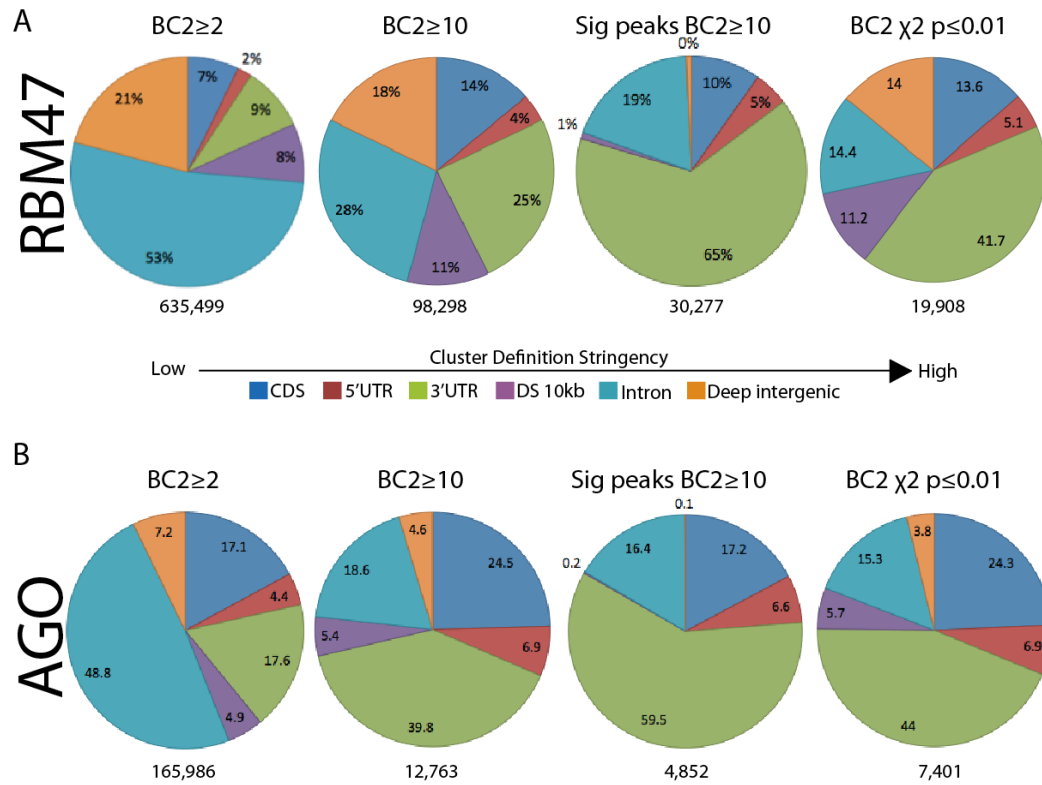


FIGURE 5.8: Genomic distribution of RBM47 clusters defined with increasing stringency indicates robust and reproducible binding predominantly in 3'UTR regions. **A:** Unique RBM47 tags from each replicate were clustered requiring a minimum of 3nt overlap between adjacent tags, then filtered with increasingly stringent criteria to reveal the most robust binding locations within the genome. At the highest level of stringency (Chi squared test of robustness), the majority of RBM47 binding is found in 3'UTR regions plus the 10kb region immediately downstream of annotated polyA sites (DS 10kb). **B:** Filtering of the Ago-mRNA HITS-CLIP data set from 831BrM2 cells shows a similar shift in distribution of binding from predominantly intronic to 3'UTR with increasing stringency of cluster definition. The number of clusters retained with each definition is shown below the pie chart.

stringency of cluster definition for both RBM47 and AGO HITS-CLIP clusters lead to a decrease in the percentage of tags mapping to introns and deep intergenic regions, retention of coding sequence clusters and enrichment for 3'UTR binding. A similar trend towards enrichment in 3'UTR binding through selective loss of intronic binding has been reported for the distribution of neuronal Elavl HITS-CLIP clusters with increasingly stringent biological complexity requirements (Ince-Dunn et al., 2012).

Motif enrichment in RBM47 Binding sites

Robust HITS-CLIP binding sites have consistently been found to be enriched in validated binding motifs of the RNA binding protein interrogated. For example, the YCAY binding motif, which is necessary and sufficient to confer Nova-dependent regulation on alternatively spliced transcripts (Dredge and Darnell, 2003, Dredge et al., 2005, Jensen et al., 2000) is highly enriched in Nova HITS-CLIP clusters (Licatalosi et al., 2008). Furthermore, binding motifs can be identified at single nucleotide resolution through analysis of crosslink-induced mutations (CIMS) in HITS-CLIP data, with reproducible deletions in robustly bound sites defining the crosslinked nucleotide for multiple RNA binding proteins (Zhang and Darnell, 2011).

To identify the RBM47 binding motif from HITS-CLIP, word enrichment was assessed in robust RBM47-binding sites. Sequences corresponding to the entire footprint of $BC2 \geq 10$ tags clusters, the 50nt

region centered on significant peaks (BC2 sigpeak \geq 10tags), and the 20nt region centered on RBM47-deletion CIMS (FDR \leq 0.01 and \geq 5 deletion harboring tags, 537) were used for motif analysis. For each subset of RBM47 binding sites a matching set of randomly generated background sequences were generated with the same base composition and nucleotide length. MEME analysis (Bailey and Elkan, 1994) was carried out using the smaller CIMS data set.

Motif enrichment for each increasingly stringent subset of RBM47 binding sites consistently returned U₁₀ as the most enriched 10mer (Figure 5.9A-C). Additionally, this sequence was the most enriched motif identified by MEME analysis of robust CIMS sites (Figure 5.9D). Analysis of nucleotide bias around the CIMS site revealed an average base composition of 27.5% A, 18.6% C, 19.9% G and 33.9% U, with 51.8% of all CIMS occurring at a U residue (Figure 5.9E). Together this would strongly suggest that RBM47 preferentially binds predominantly U-stretches or AU-rich sequences (AREs) in target RNA.

RBM47 re-expression in 831BrM2 cells leads to steady state levels changes of RBM47-bound transcripts

To gain insight into the potential functional role of RBM47 RNA binding, transcript level changes associated with RBM47 re-expression in 831BrM2 cells were assessed. RNAseq libraries were prepared from one dish each of Tet-On and RBM47-WT cells treated with 1ng/ml dox for three days prepared at the same time as those used for HITS-CLIP.

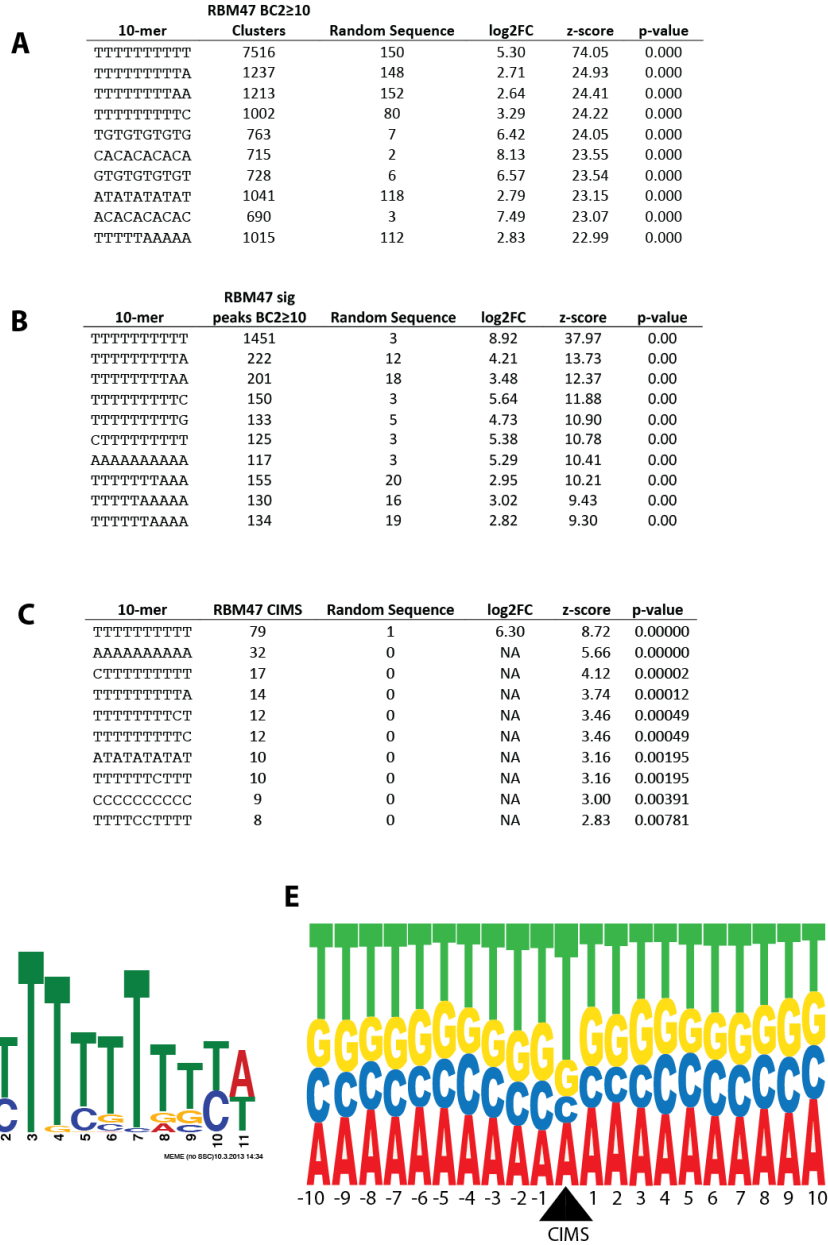


FIGURE 5.9: Motif enrichment in RBM47 binding sites with increased stringency of binding site definition. Word enrichment in sequences corresponding to all BC2≥10tags clusters (**A**), a 50nt footprint centered on significant peaks BC2≥10tags (**B**), or in a 20nt footprint centered on crosslink-induced deletions (CIMS FDR≤0.001 m≥5, **C**). All RBM47-bound sequences were compared to randomly generated sequences of equal number, length and base composition. **D**: MEME analysis of the 20nt CIMS footprint revealed significant enrichment for a polyT motif (63/537 sites p=3.0e⁻¹⁹). **E**: Sequence composition anchored at RBM47 CIMS reveals AU-rich sequence with the crosslink site most frequently mapped to uridine.

~169million 100nt paired-end RNAseq reads were mapped to the human genome (hg18), ~51million from the Tet-On control cells and ~118million from RBM47-WT cells. Fold change and associated p-value for each transcript was calculated based on a binomial distribution of reads per transcripts, and an FDR calculated using the Benjamini-Hochberg correction for multiple hypothesis testing. Transcripts were considered “expressed” given a minimum of two RNAseq reads per transcript in both experimental conditions (16,466 genes with ≥ 2 RNAseq reads per transcript).

Plotting read count per expressed gene from Tet-On cells against those from RBM47-WT revealed a high correlation between steady state transcript levels ($R^2=0.67$) suggesting that global transcription was not altered by the expression of RBM47 (Figure 5.10). To identify transcripts with robust RBM47-dependent changes in expression level, genes were filtered by fold change $\geq |2|$ and $FDR \leq 0.01$. A highly conservative filter was then applied to identify genes with expression level changes greater than two standard deviations from the linear regression line to isolate those transcripts with robust total RNAseq counts in an attempt to reduce noise inherent in the interpretation of expression level changes derived from single replicate RNAseq libraries.

39 upregulated and 64 downregulated targets were identified, all of which had evidence of RBM47 binding at a minimal definition of one BC2 ≥ 2 tags cluster. Increasing the requirement for RBM47 binding to ≥ 100

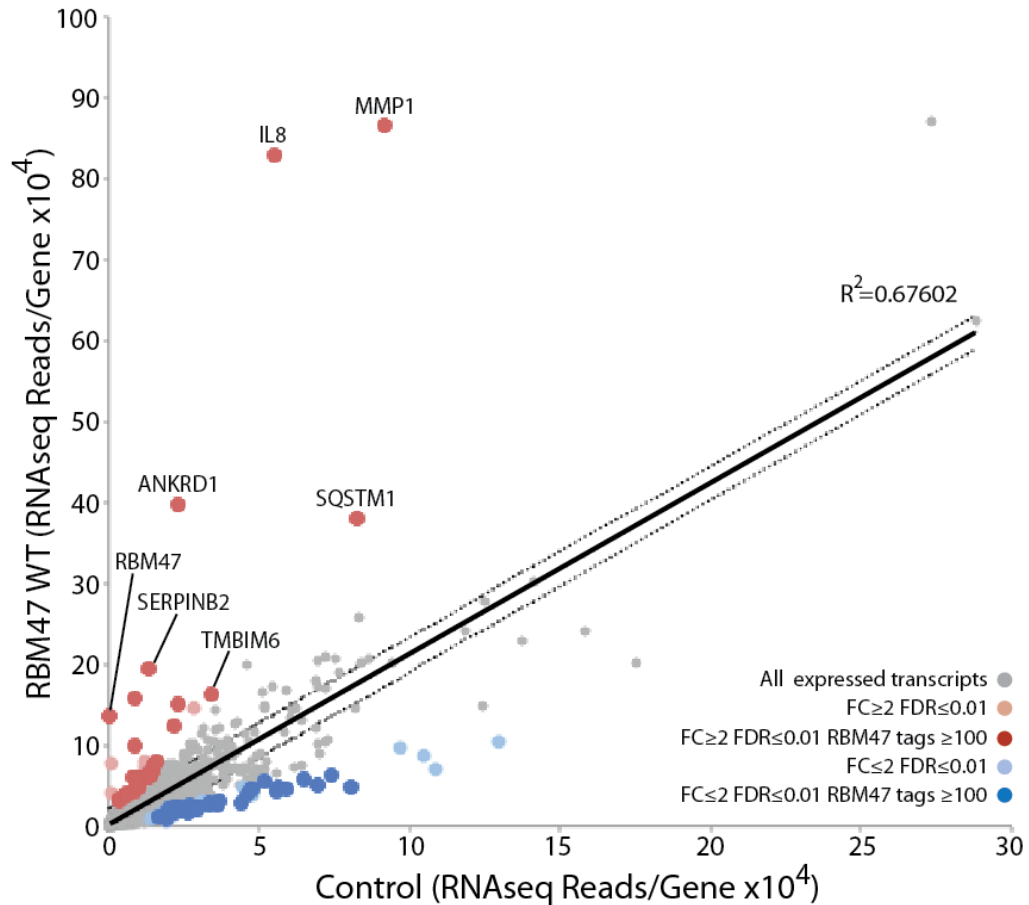


FIGURE 5.10: RNAseq analysis reveals RBM47-dependent steady state transcript level changes. The transcript levels of all expressed genes (greater than two RNAseq reads in both samples, grey) correlate well between Tet-On control and RBM47-WT expressing 831BrM2 cells. To identify differentially expressed targets, genes were selected that showed fold changes in RNAseq reads of ≥ 2 (light red) or ≤ 2 (light blue), $FDR \leq 0.01$ and level changes greater than two standard deviations (dotted line) from the linear regression (black line). Genes were then screened for direct RBM47 regulation by the requiring ≥ 100 total RBM47 HITS-CLIP tags in clusters ($BC2 \geq 2$ tags) per upregulated (dark red) and downregulated (dark blue) transcript. R^2 =Pearson product moment correlation coefficient.

total tags in BC2 \geq 2tags clusters further refined the set of transcripts to 32 RBM47-bound and upregulated genes, and 46 RBM47-bound and downregulated genes (Figure 5.10). RBM47 itself passed the threshold for upregulation with a fold change of 240 and FDR $4.55e^{-15}$. Interestingly 2,060 RBM47 tags mapped to the RBM47 transcript, suggesting potential for RBM47 auto-regulation. The RBM47-bound up- and downregulated genes are listed in Table 5.1 and 5.2 respectively. The DAVID Bioinformatics database was used to analyze the gene ontology (GO) terms associated with RBM47 bound and regulated transcripts but no enrichment in terms was found (FDR \leq 0.05).

For those genes positively regulated by RBM47, RNAseq transcript level fold change was positively correlated to the total number of RBM47 CLIP tags in BC2 \geq 2tags clusters bound to the transcript ($R^2=0.42$, Figure 5.11A) while RBM47-downregulated genes had no correlation between the number of RBM47 tags and transcript fold change. The majority of RBM47 binding in upregulated transcripts occurred in 3'UTRs (~85%), with binding in downregulated transcripts more evenly distributed between CDS, intron and 3'UTR (Figure 5.11B). RNAseq reads per gene in RBM47-WT cells were plotted against the total number of tags in BC2 \geq 2tags clusters in the gene (Figure 5.12). No overall correlation was found between the abundance of transcript and RBM47 binding ($R^2=0.06$) indicating RBM47 does not indiscriminately bind all upregulated transcripts.

TABLE 5.1: RBM47-bound transcripts showing RBM47-dependent increase in steady state transcript level by RNAseq

Gene ID	Gene	Fold Change	FDR	RBM47 CLIP tags
22943	DKK1	8.77	7.94E-15	20505
3576	IL8	6.93	1.57E-14	17390
1490	CTGF	2.66	4.58E-15	10543
27063	ANKRD1	8.01	1.14E-14	7590
7009	TMBIM6	2.23	8.43E-15	7219
5552	SRGN	2.60	5.52E-15	7001
6781	STC1	2.24	4.56E-15	5005
4312	MMP1	4.37	1.72E-14	4478
23204	ARL6IP1	2.50	3.61E-15	3925
55161	TMEM33	3.92	3.15E-15	3552
5055	SERPINB2	6.96	6.97E-15	3038
481	ATP1B1	2.73	5.53E-15	2904
1075	CTSC	3.55	4.11E-15	2445
3552	IL1A	2.30	4.57E-15	2152
54502	RBM47	239.03	4.55E-15	2060
7056	THBD	2.18	3.63E-15	1868
1974	EIF4A2	2.23	5.05E-15	1476
8878	SQSTM1	2.12	1.38E-14	1467
57823	SLAMF7	4.29	2.26E-15	1324
5954	RCN1	2.30	3.64E-15	1312
2919	CXCL1	5.33	5.52E-15	1132
378	ARF4	2.02	5.05E-15	971
9741	LAPTM4A	2.17	4.56E-15	914
3383	ICAM1	2.23	4.11E-15	729
5806	PTX3	3.21	4.10E-15	645
58505	OSTC	2.82	4.55E-15	556
55969	C20orf24	3.27	4.08E-15	527
10627	MYL12A	2.06	3.17E-15	367
9550	ATP6V1G1	3.15	3.62E-15	214
6520	SLC3A2	3.00	6.97E-15	192
3553	IL1B	2.82	4.57E-15	144
11009	IL24	2.32	5.51E-15	122

TABLE 5.2: RBM47-bound transcripts showing RBM47-dependent decrease in steady state transcript level by RNAseq.

Gene ID	Gene	Fold Change	FDR	RBM47 CLIP tags
667	DST	-2.54	1.35E-98	10437
6711	SPTBN1	-2.50	1.58E-99	5803
23499	MACF1	-2.60	7.54E-100	4672
3655	ITGA6	-2.22	3.02E-99	4401
4853	NOTCH2	-2.94	2.23E-99	3766
23215	PRRC2C	-2.14	7.92E-100	3358
79026	AHNAK	-2.54	5.35E-100	2991
64750	SMURF2	-2.08	5.43E-100	2806
8829	NRP1	-2.59	1.22E-99	1659
7837	PXDN	-2.62	1.36E-99	1590
2317	FLNB	-2.26	4.21E-99	1590
3673	ITGA2	-2.76	3.00E-99	1385
4134	MAP4	-2.71	2.59E-99	1227
10075	HUWE1	-3.15	9.90E-100	1118
3069	HDLBP	-2.04	3.45E-99	920
4898	NRD1	-2.23	2.21E-99	811
51520	LARS	-2.03	6.59E-100	793
3376	IARS	-3.01	3.18E-99	641
5591	PRKDC	-2.79	1.89E-99	525
6574	SLC20A1	-2.53	1.64E-99	520
54443	ANLN	-2.69	6.47E-100	439
7468	WHSC1	-3.20	1.40E-99	438
9448	MAP4K4	-2.37	1.13E-99	435
3949	LDLR	-3.25	2.77E-99	418
3609	ILF3	-2.42	3.07E-99	412
7153	TOP2A	-2.77	1.46E-99	406
23524	SRRM2	-3.51	7.49E-100	392
81	ACTN4	-2.85	9.69E-98	356
7175	TPR	-2.52	1.45E-99	328
9793	CKAP5	-2.85	1.04E-99	281
8570	KHSRP	-2.42	1.27E-99	270
1305	COL13A1	-2.50	7.48E-99	257
11100	HNRNPUL1	-2.07	8.79E-100	247
3371	TNC	-3.20	3.19E-99	246
2335	FN1	-3.97	4.17E-99	245
5339	PLEC	-3.64	1.99E-99	230
5327	PLAT	-2.45	2.01E-99	214
7094	TLN1	-2.85	1.47E-99	194
10594	PRPF8	-2.90	9.25E-100	186
1063	CENPF	-4.36	8.81E-99	183
5214	PFKP	-2.37	2.08E-99	154
3675	ITGA3	-3.02	3.003E-99	150
4017	LOXL2	-2.13	2.68E-99	130
1786	DNMT1	-2.46	5.84E-99	112
3710	ITPR3	-3.00	2.94E-99	108
7916	PRRC2A	-2.72	1.35E-99	106

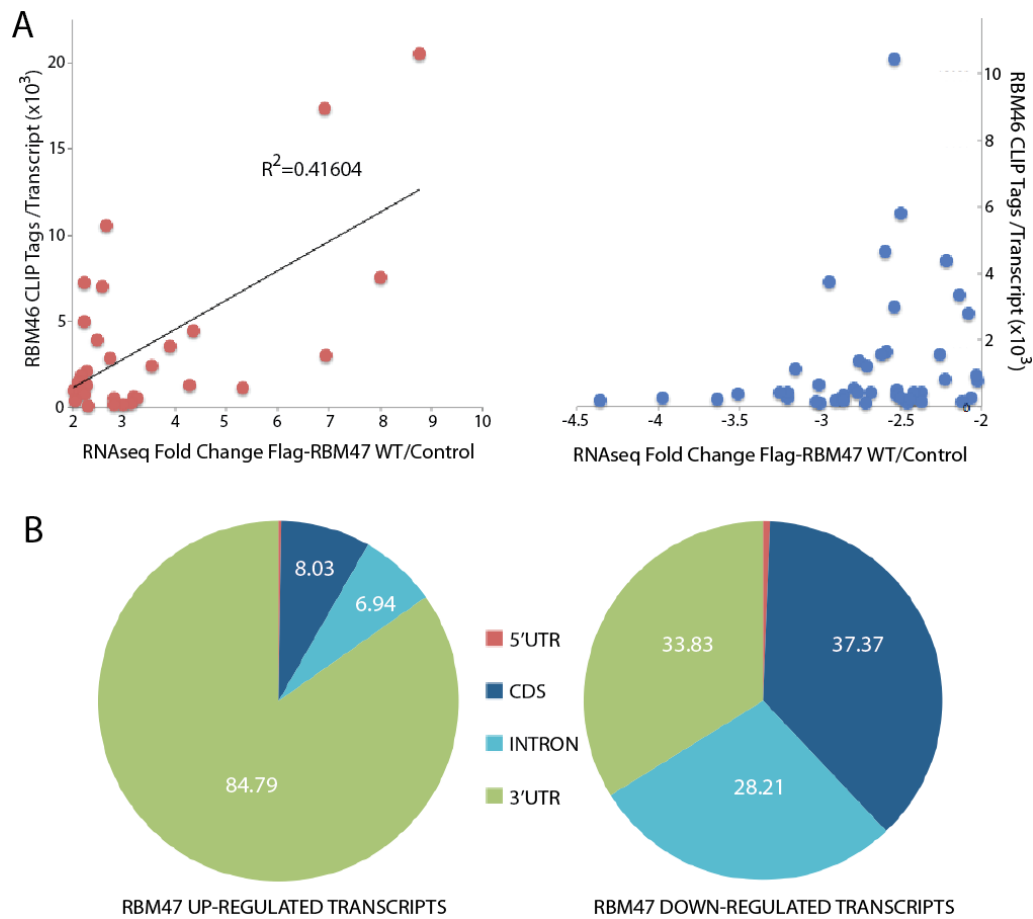


FIGURE 5.11: RBM47 binding in 3'UTRs positively correlates with transcript up regulation. **A:** RBM47 CLIP tags per transcript were plotted against RBM47-dependent transcript fold change for upregulated (red) transcripts, revealing a positive correlation between the RBM47 CLIP tags in clusters and RBM47-dependent fold induction. No correlation between number of binding events and fold change was seen for RBM47-bound downregulated targets (blue). **B:** The percentage of RBM47 CLIP tags in $BC2 \geq 2$ tags clusters in each region of RBM47-bound and regulated transcripts were calculated and averaged for up- and downregulated genes. The ~85% of RBM47 binding associated with RBM47-dependent transcript upregulation occurs in 3'UTRs, while binding is more even dispersed along RBM47-bound and downregulated transcripts.

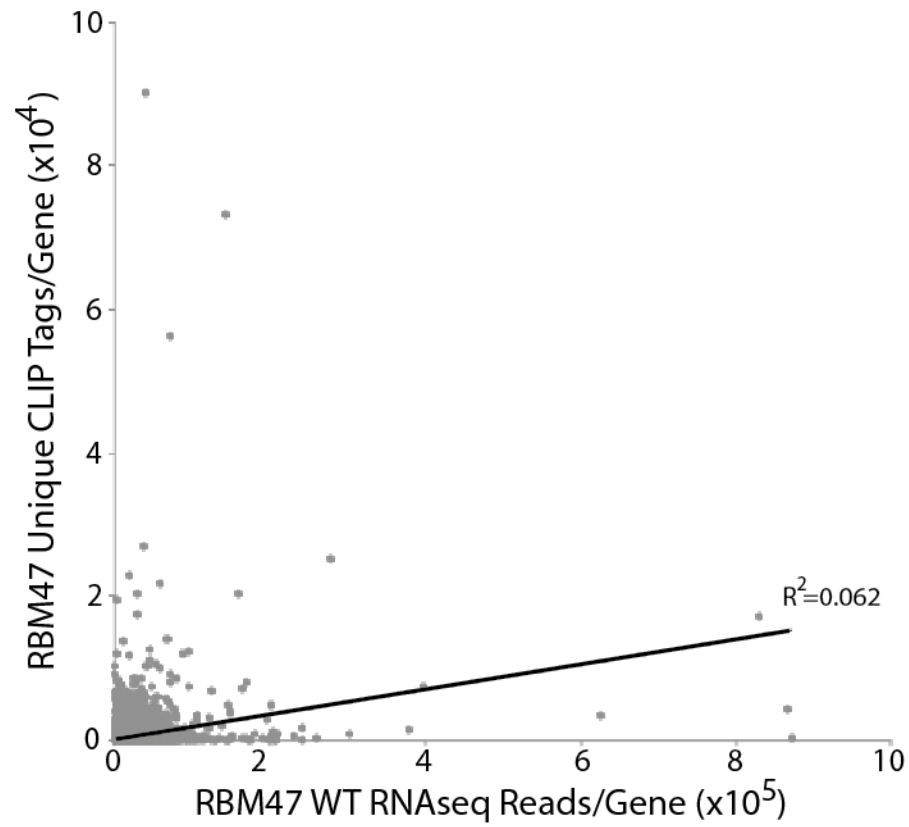


FIGURE 5.12: RBM47 binding does not correlate with transcript abundance. RNAseq reads/transcript in RBM47-WT cells are plotted against RBM47 unique CLIP tags in BC2 \geq 2tags clusters/transcript.

Independent validation of RBM47 mediated transcript level changes

Validation of RBM47-mediated transcript level change was carried out using a clonal cell line derived from the original RBM47 WT cell line used for HITS-CLIP and RNAseq analysis. 831BrM2 Flag-RBM47 WT clone#10 (RBM47-WT#10) has tightly regulated inducible RBM47 expression compared to the original non-clonal line, and was used in the *in vivo* metastasis suppression assays outlined in Figure 5.2. Cells were treated for 3 days with 10ng/ml dox and RBM47 induction validated by Q-PCR and Western blot (Figure 5.13) with the original non-clonal 831BrM2 Tet-On cell line used as a negative control. Transcript level of bound targets was assessed by Q-PCR comparing RBM47-WT#10 to Tet-On control, and by comparing MDA-231 (RBM47 positive) to 831BrM2 (RBM47 negative, see Figure 5.1) to determine whether presence of endogenous RBM47 recapitulated exogenous RBM47-dependent transcript level changes.

Comparison of RBM47 bound and regulated targets to previously identified 831BrM2 signature expression changes

An extensive reference data set of genes whose expression levels change in 831BrM2 relative to the less metastatic MDA-231 cells (Bos et al., 2009) was compared to the genes listed in Tables 5.1 and 5.2. None of the genes identified as upregulated RBM47 targets had previously identified as genes downregulated in 831BrM2 compared with MDA-231. A subset of the RBM47-bound and upregulated transcripts had previously been identified as being upregulated in 831BrM2, including CTGF, SRGN,

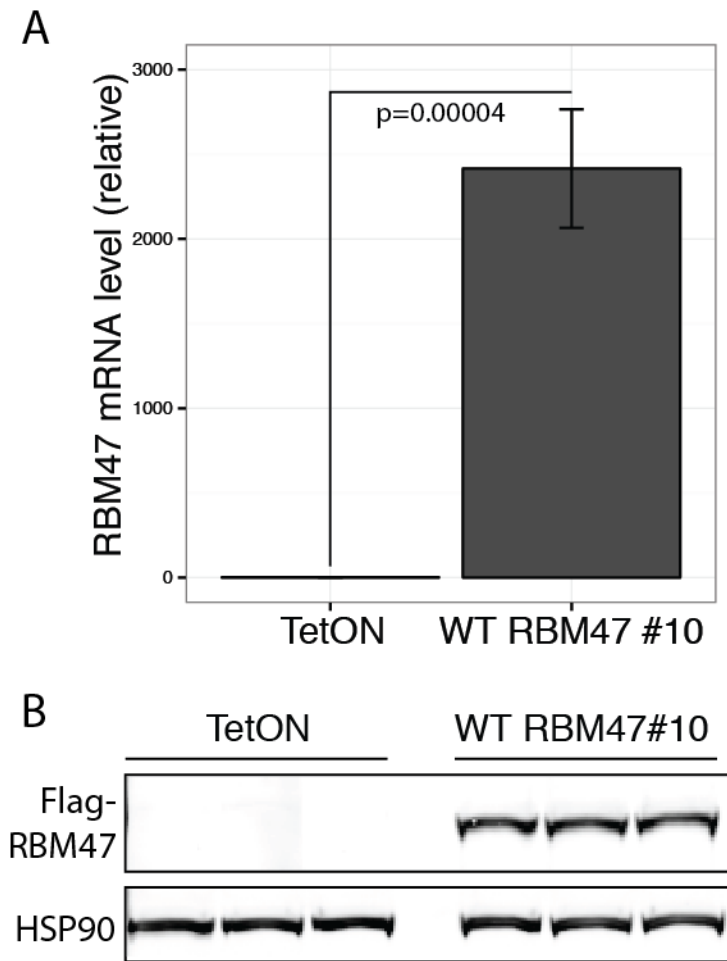


FIGURE 5.13: Verification of dox-inducible Flag-RBM47 expression in the clonal population of 831BrM2 Tet-On Flag-RBM47 WT used for validation assays. A: Total RNA was prepared in triplicate from Tet-On control and WT-RBM47#10 cells treated with 10ng/ml dox for three days. Q-PCR for RBM47 revealed greater than 2000-fold induction of RBM47 transcripts level (CT values in Tet-On cells were >30). Expression levels were normalized to TBP, p-value calculated by t-test. **B:** Protein lysates were prepared in triplicate from dox-treated cells. Flag-RBM47 protein expression is seen only in WT-RBM47#10 with HSP90 shown as a loading control.

STC1, MMP1, SERPINB2, CTSC, IL1A and IL1B (Bos et al., 2009), suggesting that while RBM47 can positively regulate the levels of multiple genes whose up regulation is associated with gain of brain metastatic potential, loss of positive RBM47 transcript regulation in 831BrM2 may be compensated for by some other undetermined transcriptional or post-transcriptional mechanism.

RBM47 binding to SERPINB2 (serpin peptidase inhibitor, clade B, member 2) is shown in Figure 5.14A (black) with the corresponding normalized RNAseq abundance data from Tet-On (red) and RBM47-WT (green, RNAseq reads/million in that library). RBM47-dependent increase in SERPINB2 transcript level is evident in the RNAseq data, and was independently verified by Q-PCR (Figure 5.14B). The expression level of SERPINB2 in MDA-231 and 831BrM2 cells was found to be consistent with previous reports, with significant increase in transcript level in brain metastatic cells (Figure 5.14C). Transcript levels of a second RBM47-bound target, CTGF (connective tissue growth factor), were found to correlate with RBM47 expression in both the RBM41-WT#10 cells and, unexpectedly, between MDA-231 and 831BrM2 contrary to previously reported expression level changes (Figure 5.15).

Comparing the published lists of transcripts upregulated in brain metastatic cells with those identified as RBM47-bound and downregulated revealed only one common gene, COL13A1. However this RBM47-dependent negative regulation did not validate by Q-PCR (data not

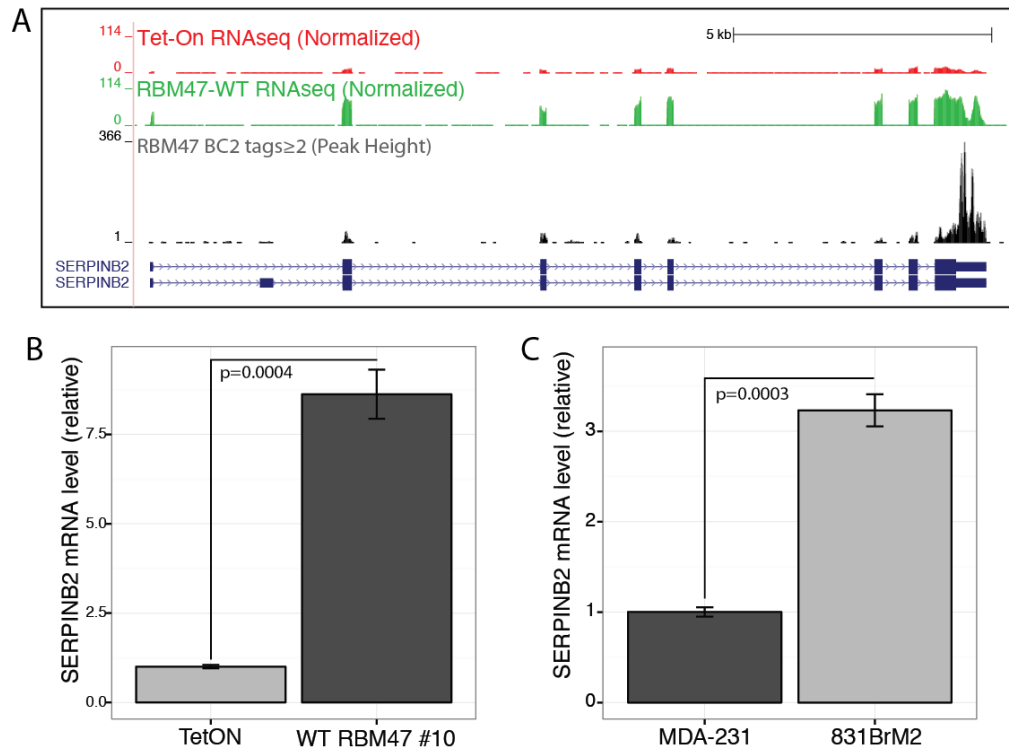


FIGURE 5.14: SERPINB2 transcript levels are positively regulated by re-expression of RBM47 in 831BrM2. **A:** The SERPINB2 gene is shown overlaid with Flag-RBM47 HITS-CLIP data (black, y-axis= number of overlapping tags in BC2 \geq 2tags clusters). SERPINB2 is bound by RBM47 predominantly its 3'UTR, with low levels of binding occurring in clusters on CDS exons and in introns. Shown above the HITS-CLIP binding map are normalized RNAseq profiles from Tet-On control (red) and RBM47-WT cells (green, y-axis=tags/million reads), with clear increase in SERPINB2 expression in the presence of RBM47 (fold change=6.96, FDR=6.97e⁻¹⁵). **B:** SERPINB2 expression levels were increased in WT RBM47#10 when compared to similarly dox-treated empty vector Tet-On control cells (fold change=8.62, p=0.0004). **C:** SERPINB2 expression levels in MDA-231 (RBM47-positive) and 831BrM2 (RBM-47 negative) concur with previously published reports of up regulation in brain metastatic cells (fold change=3.23, p=0.0003). RNA prepared from three independently treated cells lines, expression levels normalized to TBP, p-values calculated by t-test.

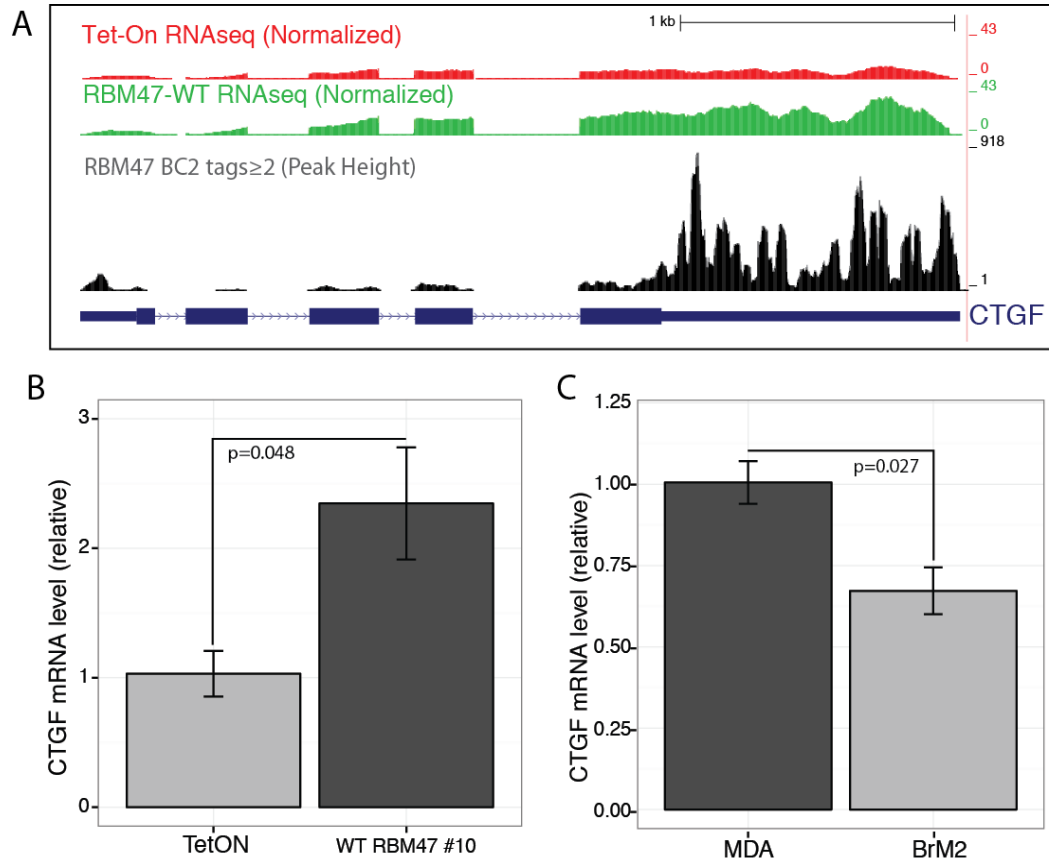


FIGURE 5.15: CTGF transcript levels correlate with RBM47 expression. **A:** The CTGF gene overlaid with Flag-RBM47 HITS-CLIP data and normalized RNAseq profiles from Tet-On control (red) and RBM47-WT (green) as in Figure 5.14A (Fold change=2.66, FDR=4.58e⁻¹⁵). **B:** Up regulation of CTGF in RBM47 expressing brain metastatic cells was confirmed by Q-PCR (Fold change=2.35, p=0.048). **C:** CTGF expression levels in MDA-231 and 831BrM2 correlate with endogenous RBM47 expression (fold change=0.67, p=0.027). RNA prepared from three independently treated cells lines, expression levels normalized to TBP, p-values calculated by t-test.

shown). Q-PCR validation was attempted for several of the RBM47-bound and downregulated targets that had shown the largest predicted fold changes by RNAseq, including MACF1, CENPF/CUGBP2 and NOTCH2. All transcripts were found to be unchanged in the presence of exogenous RBM47 and when comparing MDA-231 to 831BrM2 cells (data not shown). Closer inspection of the RNAseq data for downregulated transcripts revealed an anomaly in the distribution of RNAseq reads along the length of transcripts in the RBM47-WT library as illustrated in Figure 5.16 on the AHNK transcript (fold change=-2.54, FDR=5.35e⁻¹⁰⁰).

While a 3'-end bias is known to occur in polyA selected RNAseq libraries (Levin et al., 2010, Tariq et al., 2011), it is not clear when comparing single RNAseq libraries whether the apparent increased rapidity of signal decay in the 3'-5' direction in RBM47-WT cells results from technical error in library preparation, or reflects a biological consequence of RBM47 binding. RBM47 bound downregulated transcripts were found to be significantly longer than all expressed transcripts (Figure 5.17). Longer transcripts have higher probability of high RNAseq reads therefore statistical analysis of even small fold changes are more robust and have lower FDRs (Oshlack and Wakefield, 2009). Visual inspection of multiple non-RBM47 bound transcripts of similar genomic length and equal differential expression by RNAseq revealed similar RNAseq profiles suggesting that the apparent RBM47-dependent downregulation of long transcripts is an artifact.

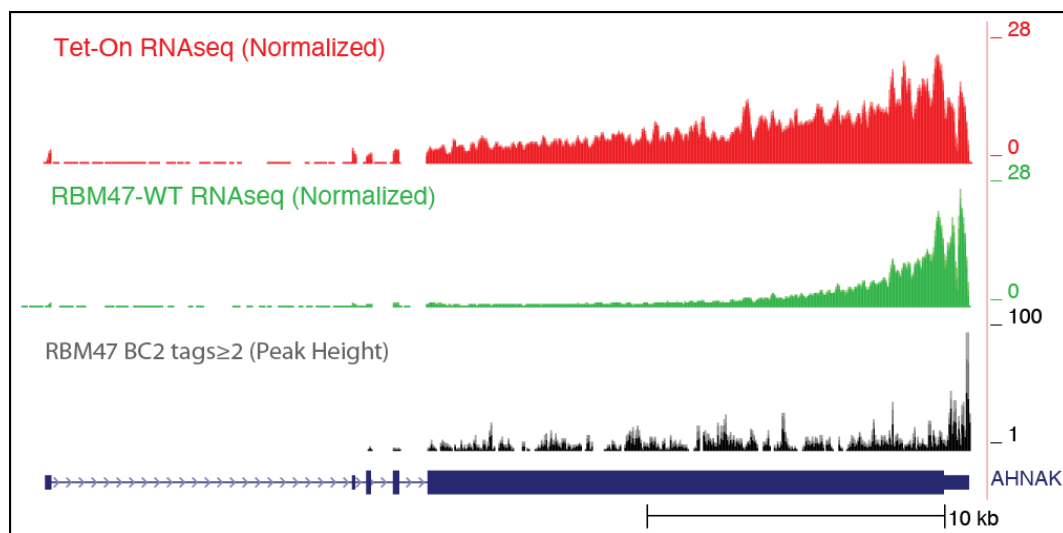


FIGURE 5.16: RBM47-bound and downregulated genes show strong 3' end bias in the RNAseq profile. The AHNK transcript (fold change=-2.54, FDR= $5.35e^{-100}$) is shown as an example of the rapid 3'-5' decay of RNAseq reads that is exaggerated in the RBM47-WT RNAseq data set compared to Tet-On control cells.

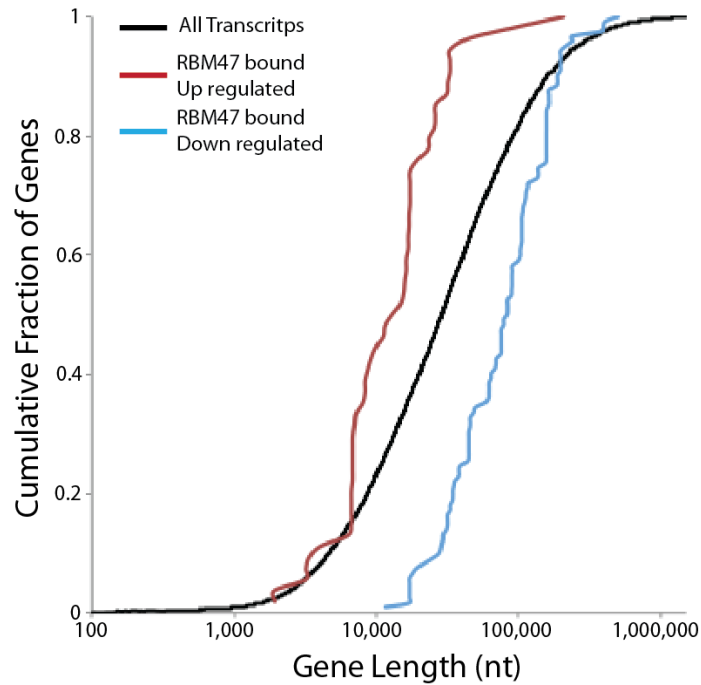


FIGURE 5.17: Cumulative distribution plots of genomic gene length of RBM47-bound genes showing RBM47 dependent transcript level changes. Genomic length of all expressed genes in the RNAseq data sets are shown in black, the 32 RBM47-bound and up regulated transcripts in red, the 46 down regulated genes in blue. RBM47-bound up regulated genes are significantly shorter than the general transcript population ($p=4.661e^{-08}$), while down regulated are significantly longer ($p<2.2e^{-16}$). Gene length was calculated using coordinates from obtained from Refseq, p-values calculated using the Kolmogorov-Smirnov test.

Further RNAseq sampling, preferably prepared using a ribo-minus protocol, needs to be carried out before any conclusions can be made about the role of RBM47 in transcript down regulation.

Validation of RBM47-mediated expression level changes in transcripts not previously identified as altered in metastasis

Ankyrin repeat domain 1 (ANKRD1) was the second most differentially expressed RBM47-bound transcript identified by RNAseq (fold change=8.01, 7590 RBM47 CLIP tags) (Figure 5.18A). The positive regulation of ANKRD1 transcript levels by RBM47 was confirmed in RBM47-WT#10 dox-treated cells (Figure 5.18B), while decrease in expression was also seen in 831BrM2 relative to MDA parental cells (Figure 5.18C). Dickkopf 1 (DKK1) was identified as both the most upregulated of RBM47-bound transcripts (fold change = 8.77) and the most highly bound RBM47-regulated transcript (20,505 RBM47 tags in BC2 \geq 2tags clusters). The majority of DKK1 RBM47 CLIP tags map to the 3'UTR (96.7%), with low level binding also evident on exons and the 5' UTR (Figure 5.19A). Q-PCR revealed ~4-fold up-regulation of DKK1 in dox-treated RBM47-WT#10 cells compared with similarly treated Tet-On controls, independently confirming the RNAseq data (Figure 5.19B). Reciprocal decrease in DKK1 transcript levels were seen in 831BrM2 compared to MDA (Figure 5.19C).

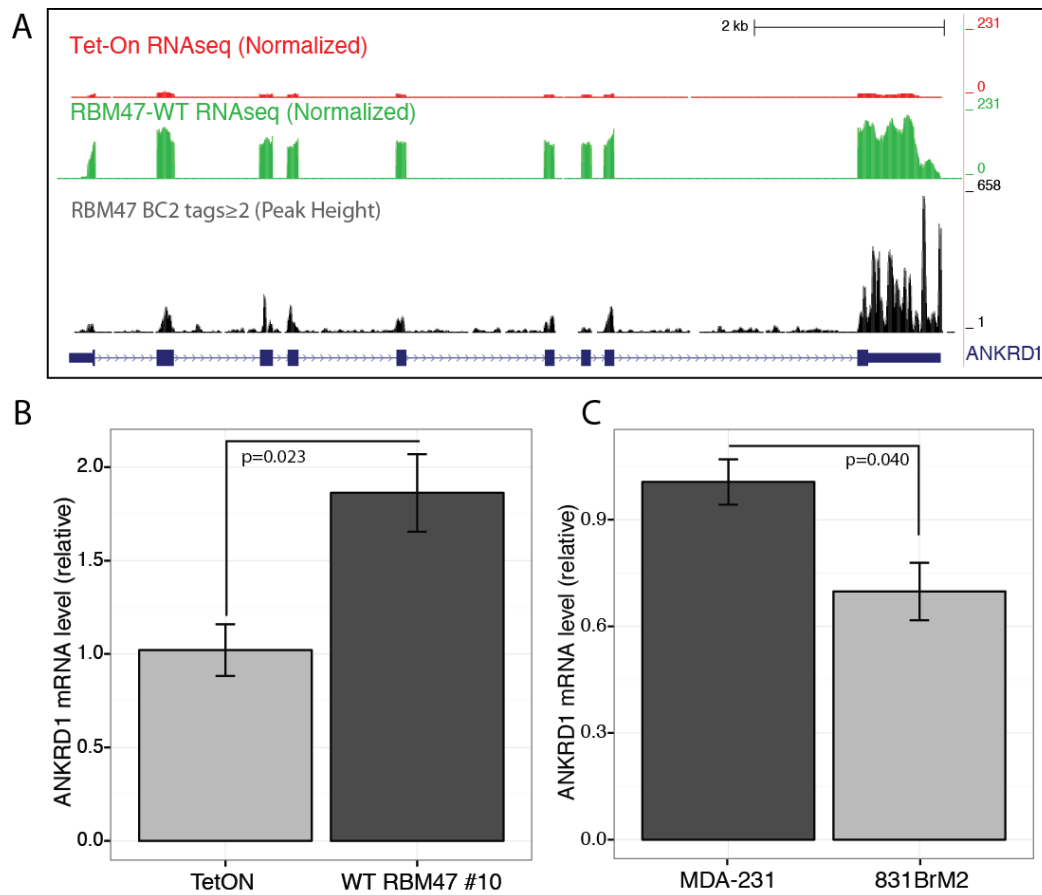


FIGURE 5.18: ANKRD1 transcript levels correlate with RBM47 expression. **A:** The ANKRD1 gene overlaid with Flag-RBM47 HITS-CLIP data (black) and normalized RNAseq profiles from Tet-On control (red) and RBM47-WT (green) (Fold change=8.01, FDR=1.14e⁻¹⁴). ANKRD1 is bound by RBM47 predominantly its 3'UTR, with lower levels of binding occurring in clusters on CDS exons and along the entire length of multiple introns. **B:** Up regulation of ANKRD1 in RBM47 expressing brain metastatic cells was confirmed by Q-PCR when compared to similarly dox-treated empty vector Tet-On control cells (Fold change=1.77, p=0.023). **C:** ANKRD1 expression levels in MDA-231 and 831BrM2 correlate with endogenous RBM47 expression (fold change=0.70, p=0.04). RNA prepared from three independently treated cells lines, expression levels normalized to TBP, p-values calculated by t-test.

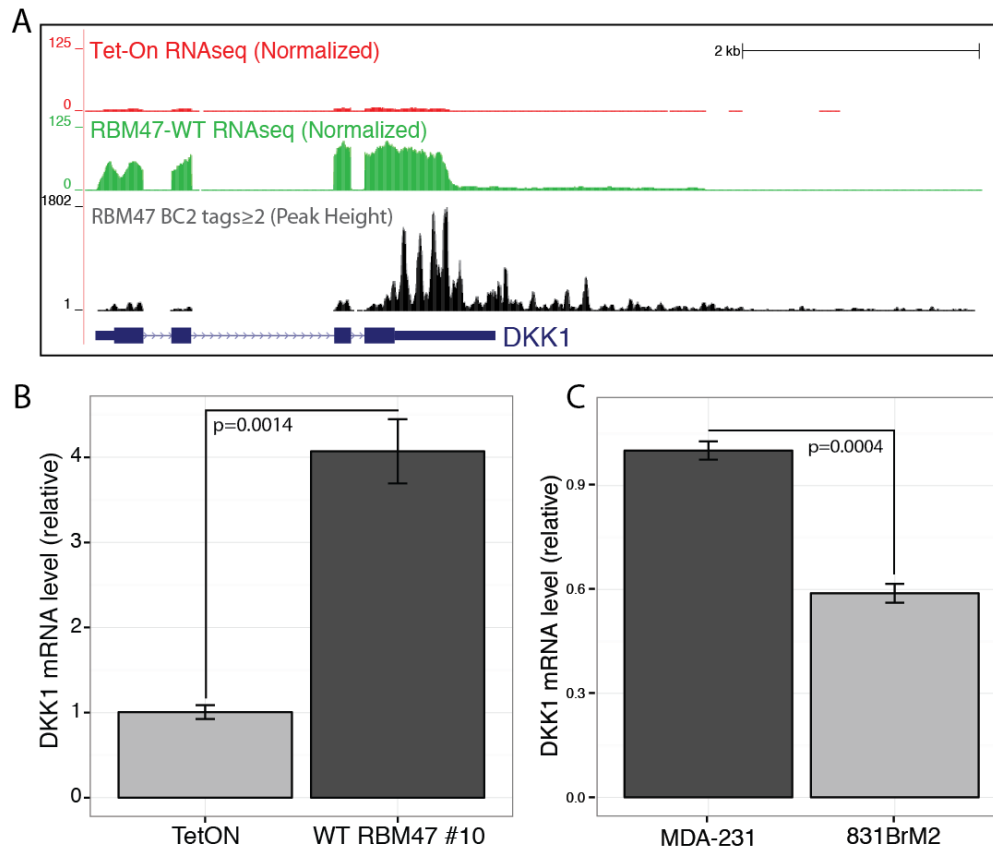


FIGURE 5.19: DKK1 transcript levels correlate with RBM47 expression.

A: The DKK1 gene overlaid with Flag-RBM47 HITS-CLIP data (black) and normalized RNAseq profiles from Tet-On control (red) and RBM47-WT cells (green) (Fold change=8.77, FDR=7.94e⁻¹⁵). Extensive binding of RBM47 is seen in an extended region beyond the annotated polyA site that is accompanied by low level RNAseq reads suggesting the presence of a minor DKK1 isoform with an extended 3'UTR. **B:** Up regulation of DKK1 in RBM47 expressing brain metastatic cells was confirmed by Q-PCR when compared to similarly dox-treated empty vector Tet-On control cells (Fold change=4.07, p=0.0014). **C:** DKK1 expression levels in MDA-231 and 831BrM2 correlate with endogenous RBM47 expression (fold change=0.59, p=0.0004). RNA prepared from three independently treated cells lines, expression levels normalized to TBP, p-values calculated by t-test.

Treatment of RBM47-WT#10 cells with α -amanitin to block transcription indicates that RBM47 binding to DKK1 leads to transcript stabilization, with the half-life of DKK1 mRNA increasing from 3.5hrs in the absence of dox-induction to 12.1hrs in the presence of RBM47 (Figure 5.20, assayed by Sakari Vanharanta, Joan Massagué laboratory, MSKCC). Given the extensive AGO binding on the DKK1 transcript identified by AGO HITS-CLIP in Chapter 4, it was hypothesized that RBM47 stabilization of the DKK1 transcripts may in part result from disruption of AGO regulation in the 3'UTR. Normalized AGO binding (tags/million unique reads in that cell line) on RBM47-dependent upregulated transcripts validated by Q-PCR (DKK1, CTGF and ANKRD1) is shown in Figure 5.21. While the normalized AGO-mRNA peaks in MDA-231 (red) and 831BrM2 (orange) are the same height on the DKK1 3'UTR (Figure 5.21A), DKK1 transcript levels are ~0.6-fold lower in 831BrM2 cells (Figure 5.19C) and therefore, on a per transcript basis, AGO binding is enhanced in RBM47-negative 831BrM2. Similarly, increase in AGO-mRNA binding is clearly seen in the 3'UTR of CTGF (Figure 5.21B) and ANKRD1 (Figure 5.21C) in 831BrM2 cells. SERPINB2 was not bound by AGO in either MDA-231 or 831BrM2 (not shown).

It has been previously demonstrated by PAR-CLIP that a positional relationship exists between exogenous tagged-HuR and tagged-AGO 3'UTR binding sites, with the peak of binding for both proteins predominantly overlapping at exactly the same nucleotide (Mukherjee et al., 2011). This observation has been independently verified using HITS-

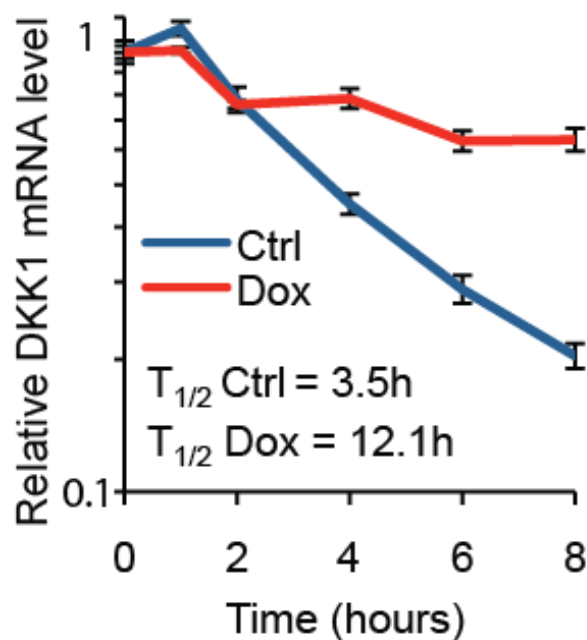


FIGURE 5.20: RBM47 expression leads to increased stability of the DKK1 transcript. RBM47-WT#10 cells were cultured in the absence (control) or presence of dox for three days before α -amanitin treatment to block RNA polymerase II transcription. The presence of RBM47 increases the half-life of the DKK1 transcript ~3.5 fold. Experiment carried out by Sakari Vanharanta, Joan Massagué laboratory, MSKCC.

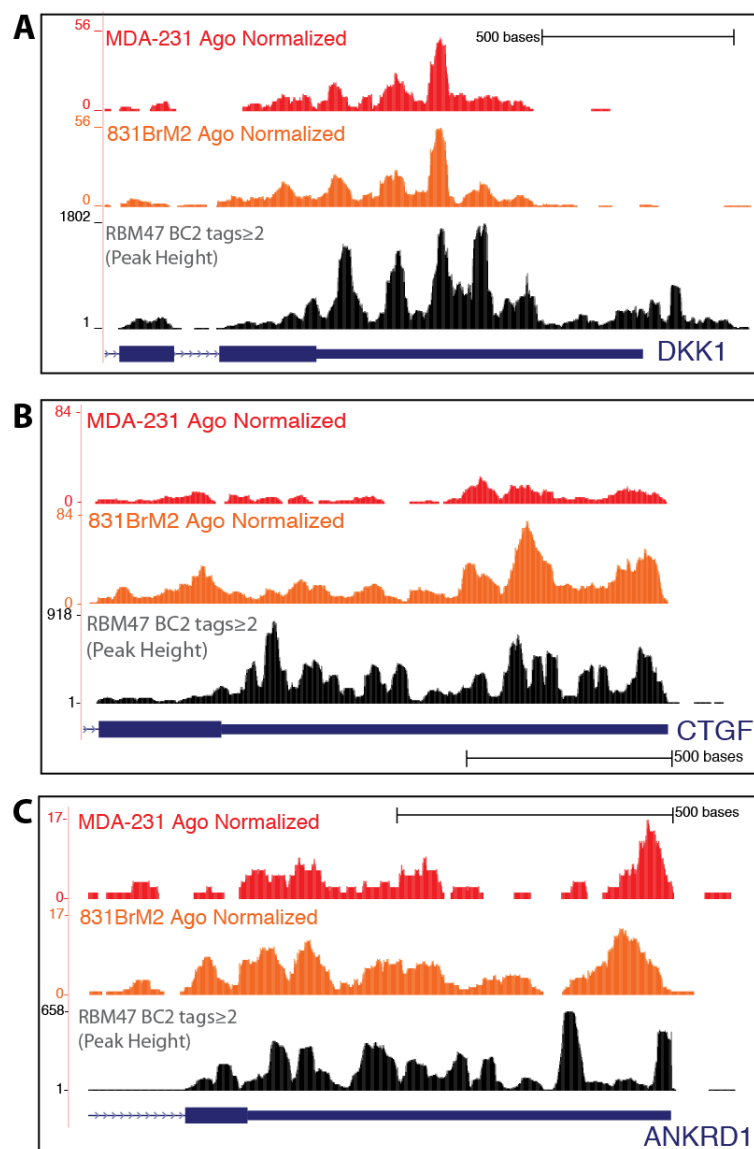


FIGURE 5.21: Normalized Ago-mRNA binding maps in MDA-231 (RBM47 positive) and 831BrM2 (RBM47 negative) on RBM47-dependent up regulated transcripts. A: Normalized Ago binding in MDA-231 (red) and 831BrM2 (orange) is shown above RBM47 binding (black) on the DKK1 transcript. DKK1 steady state transcript level is reduced in 831BrM2 cells (Figure 5.19C) so that on a per transcript basis, less Ago binding is evident in MDA-231 cells (in the absence of RBM47). A reduction in Ago binding in MDA-231 cells is evident on the CTGF 3'UTR (**B**) and ANKRD1 3'UTR (**C**).

CLIP for endogenous protein binding sites in T cells (Darnell laboratory, unpublished data), and has been suggested to either agonize or antagonize AGO binding. To see whether such a positional relationship exists between RBM47 and AGO, significant peak 3'UTR footprints of AGO in 831BrM2 cells (± 25 nt centered on peak in $BC \geq 2$ tags ≥ 10 significant peaks) were overlapped with 3'UTR significant peaks from RBM47 HITS-CLIP ($BC \geq 10$, 19,455 total). Of 2,832 AGO significant 3'UTR peaks, 43% overlapped the footprint coordinates of RBM47 significant 3'UTR peaks by a minimum of 1nt. The position of RBM47 peak binding was calculated relative to the peak of AGO binding in overlapping 3'UTR clusters (Figure 5.22). While a slight enrichment in peak binding of RBM47 is seen in regions corresponding to the ± 10 nt region of AGO peak binding, RBM47 appears to bind in a much more dispersed pattern than has been described for HuR with respect to AGO. This suggests that AGO and RBM47 do not bind exactly the same sites within 3'UTRs rather that RBM47 coats the entire footprint region in which AGO targeting occurs.

Analysis of paired-end RNAseq identifies extensive RBM47-dependent alternative splicing

Reproducible binding of RBPs in intronic regions is predictive of a role in pre-mRNA processing. To explore the relationship between RBM47 intronic binding and alternative splicing, the previously described paired end 100nt RNAseq data sets were analyzed as described for the identification of Muscleblind-like 2 (Mbnl2) regulated alternative splicing in the mouse brain (Charizanis et al., 2012). Focusing on a comprehensive

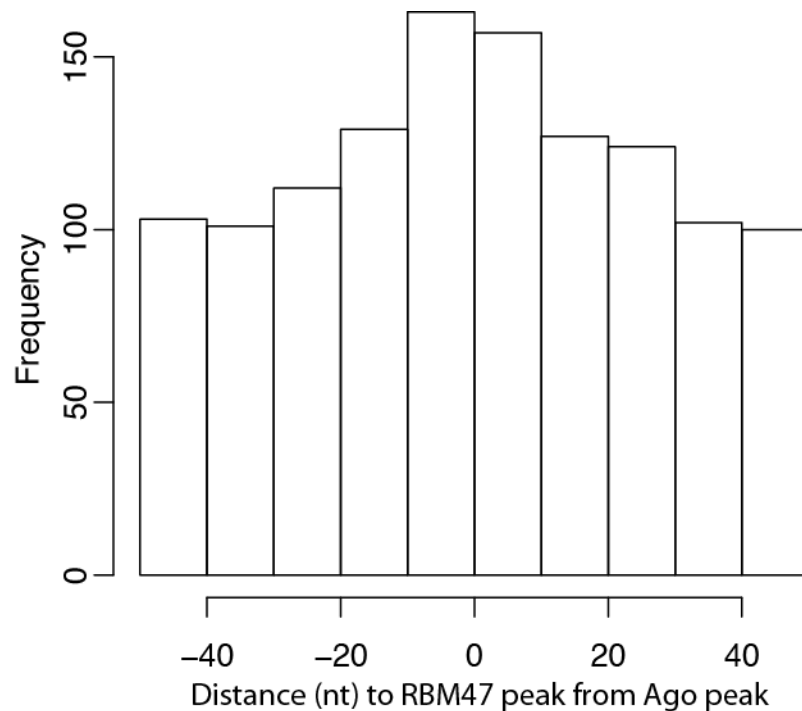


FIGURE 5.22: Histogram of the overlap of RBM47 and Ago significant peak binding in 3'UTRs. The center of all Ago significant peaks in 3'UTRs was determined and a +/-25nt footprint assigned centered on the peak coordinate. The Ago coordinates were intersected with similarly derived footprints of RBM47 significant 3'UTR peaks, resulting in 1,218 overlapping significant binding sites. The distance of the RBM47 peak from the Ago peak was calculated as is shown as a histogram. While some enrichment of overlap is seen in the -/+10nt region spanning Ago peak of binding, RBM47 peaks are more dispersed on 3'UTRs with respect to Ago binding than previously described patterns of HuR and Ago binding sites (Mukherjee et al., 2011).

data set of cassette exons derived from mRNA/expressed sequence tag data, inclusion rates (IR) were calculated for each cassette exon as outlined in Figure 5.23, and the change in inclusion rate (ΔI) upon RBM47 expression calculated. This method allows for the identification of reciprocal splicing changes between two samples with high sensitivity while normalizing for changes in RNA stability (Ule et al., 2005b). Positive ΔI indicates enhanced inclusion in the presence of RBM47, while negative ΔI indicates RBM47-dependent cassette exon skipping.

To select high confidence alternative splicing events from single RNAseq libraries, alternative splices with ≥ 10 RNAseq tags mapping to cassette exon junctions and $|\Delta I| \leq 0.2$ were considered for further analysis. This gave 669 cassette exons with RBM47 enhanced inclusion and 756 cassette exons with RBM47 enhanced exclusion. To assess direct regulation by RBM47, the genomic coordinates of the cassette exon locus, from the start of the 5' flanking exon (5'FE) across the 5' flanking intron (5'FI), cassette exon (CA), 3' flanking intron (3'FI) to the end of the 3' flanking exon (3'FE, see Figure 5.23), were cross-referenced to BC2 ≥ 2 tags RBM47-WT HITS-CLIP cluster data. Surprisingly, direct RBM47 binding was evident in the genomic locus of a high proportion of alternative splicing events showing RBM47-dependence. 480 included and 464 excluded splicing events showed evidence of at least one BC2 ≥ 2 tags cluster (Figure 5.24). The DAVID Bioinformatics database was used to analyze the gene ontology (GO) terms associated with these 944 cassette exons (located in 729 genes) using the total mRNA population present in

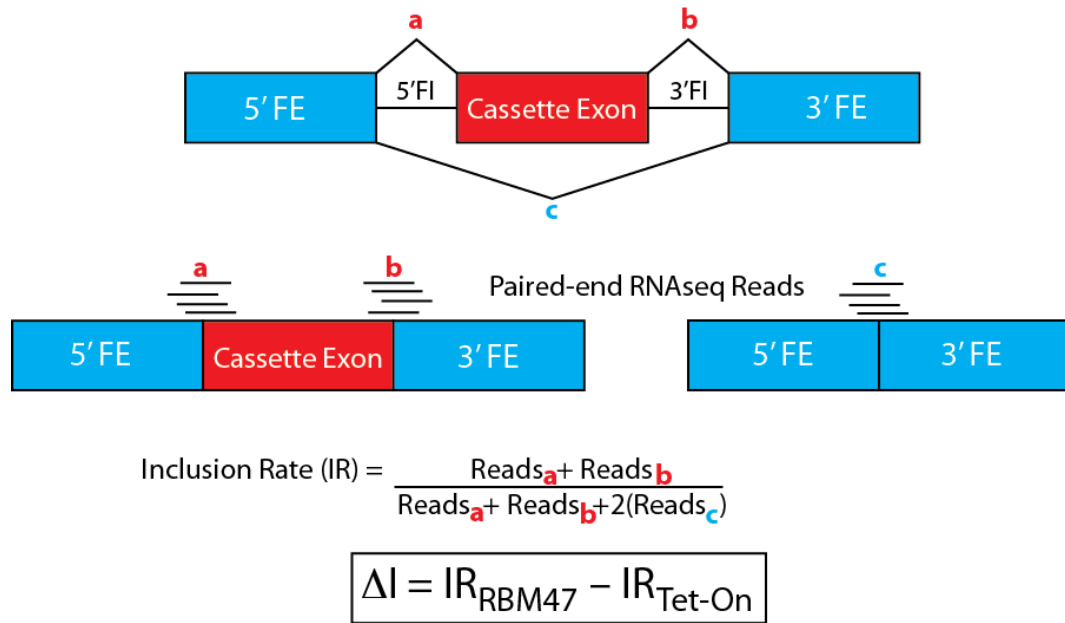


FIGURE 5.23: Calculating change in inclusion rates of alternatively spliced isoforms from paired end RNAseq data. A schematic of an alternatively spliced exon with constitutive exons in blue (5'FE - 5' flanking exon, 3'FE - 3' flanking exon) and the alternatively spliced cassette exon (CA) in red. Paired end RNAseq reads were mapped to the genome allowing for split-segment alignment to splice junctions, then further processed to infer unobserved exon junctions between aligned paired reads. The number of reads spanning constitutive-cassette (5'FE-CA and CA-3'FE) and constitutive-constitutive splice junctions (5'FE-3'FE) were used to calculate the inclusion rate (IR) of each cassette exon in a comprehensive database. The change in inclusion rate (ΔI) was calculated such that a positive ΔI indicates enhanced inclusion in the presence of RBM47, while negative ΔI indicates cassette exon skipping. A ΔI of $|0.2|$ indicates a change in inclusion rate of 20%.

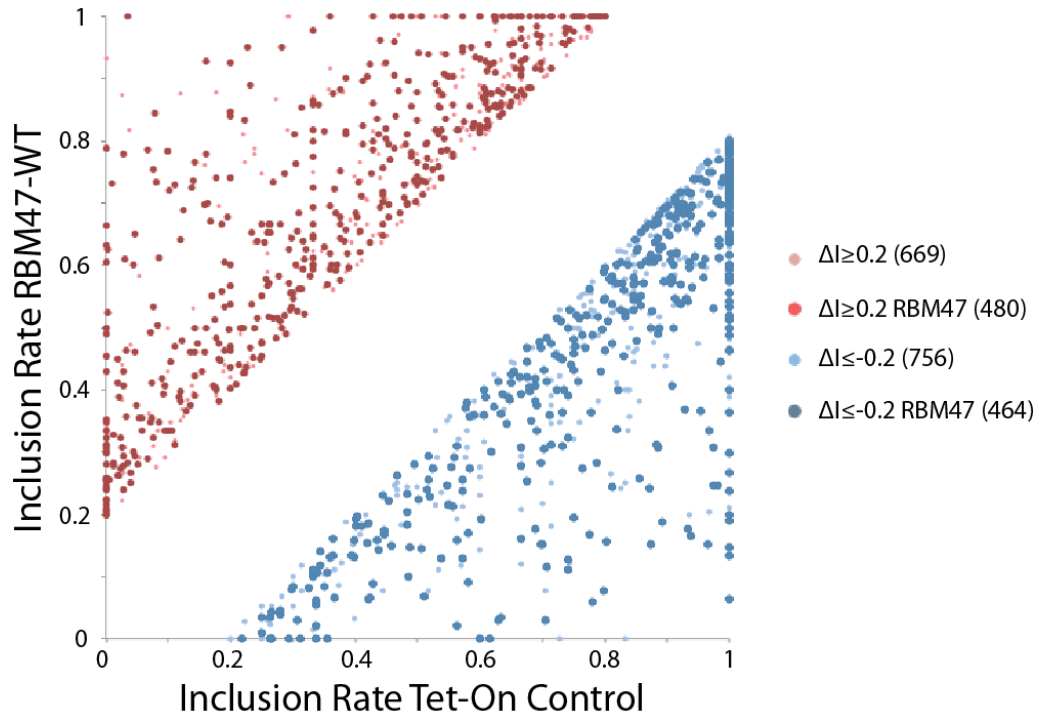


FIGURE 5.24: The majority of RBM47-dependent $\Delta I \geq |0.2|$ alternatively spliced cassette exons are bound by RBM47. Cassette exons with greater than 10 RNAseq reads mapping to splice junctions in both Tet-On control and RBM46-WT libraries and with $\Delta I \geq 0.2$ or $\Delta I \leq -0.2$ are shown in light red and light blue, respectively. Alternatively spliced exons with RBM47 binding in the genomic region spanning 5'-FE start to 3'-FE end (minimum one BC2 \geq 2tags cluster) are shown in darker shades.

831BrM cells as the background (16,466 genes with ≥ 2 RNAseq tags per transcript). Proteins encoded by RBM47-bound alternatively spliced transcripts were weakly enriched in GO terms related to the cytoskeleton and RNA binding ($p \leq 0.01$, $FDR \geq 0.01$).

The large number of RBM47-bound transcripts with RBM47-dependent altered splicing patterns indicates an extensive role for RBM47 in pre-mRNA processing. This is despite the relatively small proportion of RBM47 intronic clusters that pass the increasingly stringent cluster definition thresholds outlined in Figure 5.8. Analysis of the median tags number per $BC2 \geq 2$ tags cluster in genic regions (Figure 5.25) clearly demonstrates that individual intronic RBM47 clusters have a lower median tag count than binding sites in 5'UTR, coding sequence and 3'UTRs. Given that pre-mRNA is estimated to be present at levels 1000x lower than mRNA (Reid et al., 2009), lower levels of tag numbers in intronic binding site are to be anticipated when compared to coding sequence or UTRs. Interestingly, increasing the cut-off for the presence of RBM47-binding in alternative splice locations by total tag numbers in all $BC2 \geq 2$ tags clusters rather than total tags per individual cluster still retains high numbers of RBM47-dependent alternative splicing events. For example, a cutoff of ≥ 100 unique tags in $BC2 \geq 2$ tags clusters identifies 206 RBM47-dependent included and 196 excluded exons. Together this would suggest a pattern of dispersed, low-level binding surrounding alternative splices similar to those described for FUS (Ishigaki et al., 2012, Rogelj et al., 2012) and TDP-43 (Tollervey et al., 2011), rather than the discrete,

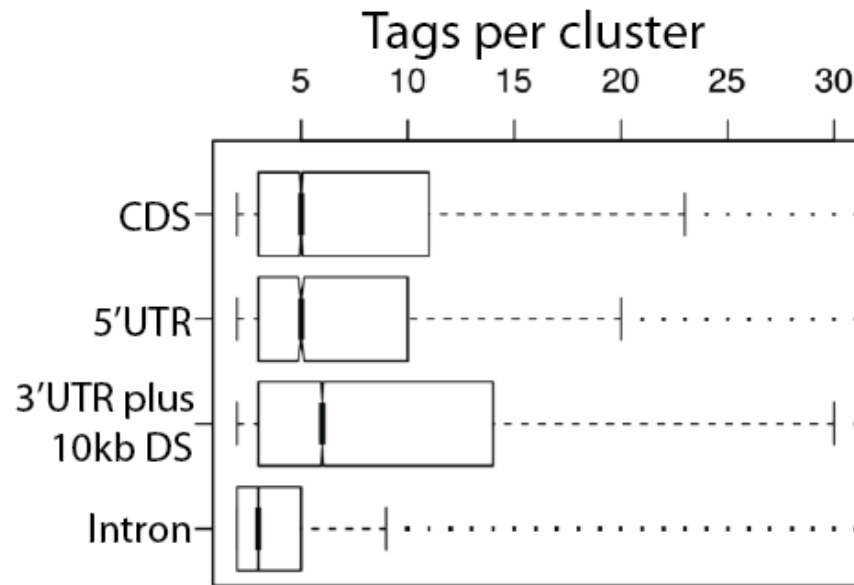


FIGURE 5.25: Distribution of RBM47 unique tag numbers in BC2 \geq 2tags clusters. The median tag number per cluster is lowest in intronic binding sites, with 95% of all intronic clusters having less than 5 tags per cluster. This explains the decrease in retained intronic clusters seen upon increase in cluster stringency definition (Figure 5.8) as applying a ≥ 10 tags per cluster threshold effectively eliminates intronic binding, retaining only outlier clusters.

robust binding sites identified by HITS-CLIP for other alternative splicing regulators such as Nova (Licatalosi et al., 2008), MBLN-1 (Wang et al., 2012), MBNL-2 (Charizanis et al., 2012) and PTBP2 (Licatalosi et al., 2012).

RBM47 binds multiple transcripts in the region of alternative splicing associated with the epithelial to mesenchymal transition (EMT)

Extensive alterations in alternative splicing patterns have been reported during the EMT transition of human mammary epithelial cells expressing tamoxifen-inducible Twist (Shapiro et al., 2011), a transcription factor essential for developmental mesoderm specification and differentiation (Qin et al., 2012). RBM47 was among several RNA binding proteins whose expression was found to decrease during EMT. Interestingly, of the 168 alternative splicing events identified by (Shapiro et al., 2011), 47 are bound by RBM47, with 36 having the same directional change in exon inclusion with respect to the presence or absence of RBM47 expression (Table 5.3).

Independent validation of RBM47-regulated splicing changes

Candidate RBM47-regulated alternative splicing events were selected for validation based on greatest observed $|\Delta I|$ and ≥ 100 RBM47 HITS-CLIP tags in BC2 ≥ 2 tags clusters in the cassette exon locus. Validation was carried out using the clonal cell line RBM47-WT#10 and 831BrM2 Tet-On cells treated for 3 days with 10ng/ml dox. Alternative splicing was assessed by semi-quantitative radioactive RT-PCR using three independent RNA preparations from each cell line. In addition,

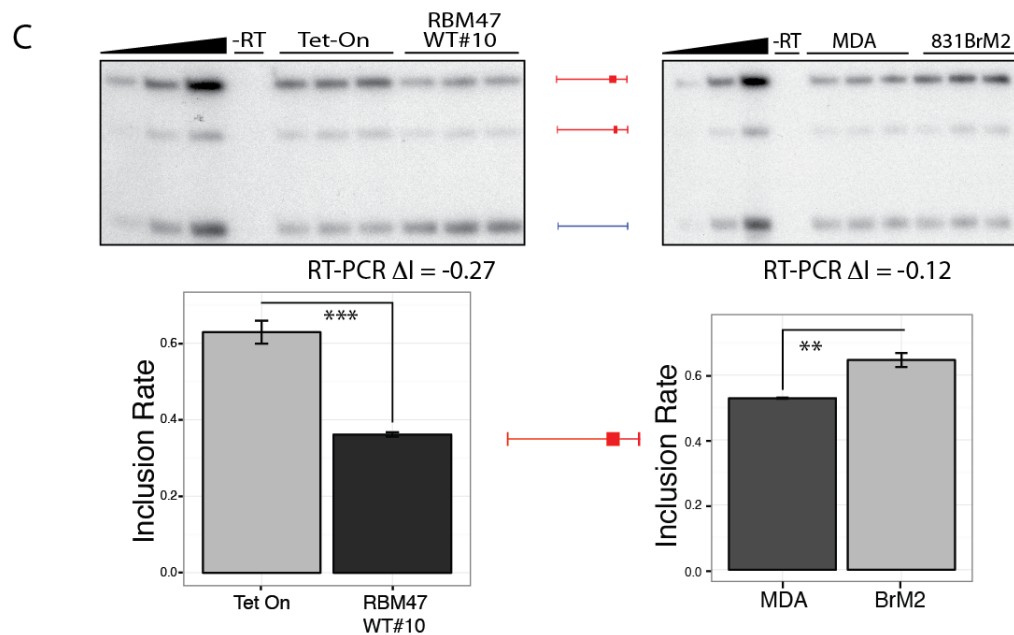
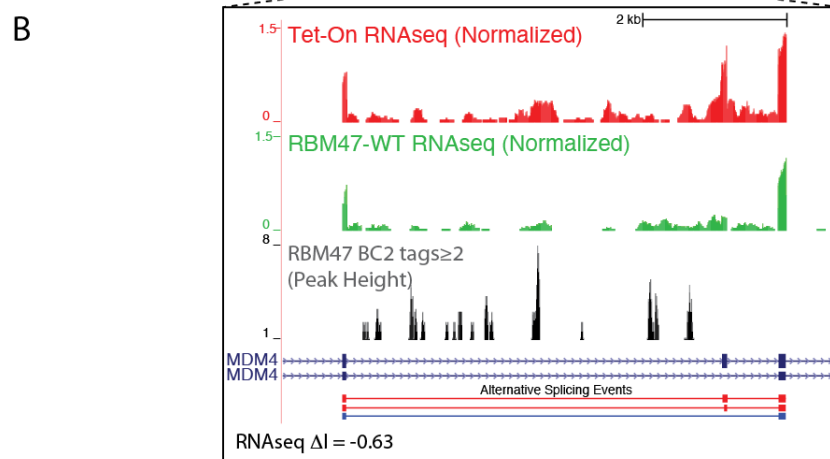
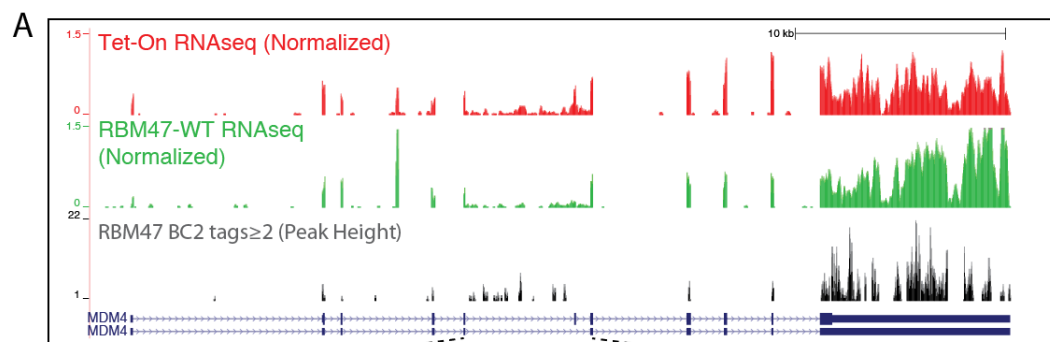
TABLE 5.3: RBM47-bound and dependent splicing changes previously identified as altered in EMT. Shown are RBM47-dependent splicing changes whose direction corresponds that seen in Shapiro et al., 2011 with respect to RBM47 expression. The number of RBM47 HITS-CLIP tags in BC2 \geq 2tags clusters in the 5'FE-3'FE genomic region of each alternative splice is given.

Gene	Chrm	5'FE Start	3'FE End	Strand	Clip tags	Delta I
ADD3	chr10	111880110	111883338	+	83	0.68
ARHGAP17	chr16	24850225	24838214	-	106	-0.27
CD44	chr11	35188087	35193036	+	547	0.26
CD44	chr11	35167957	35189571	+	3896	0.65
CLSTN1	chr1	9723900	9718529	-	20	-0.65
CTNND1	chr11	57285844	57319776	+	475	-0.27
CUGBP2	chr10	11247454	11331169	+	139	-0.29
DAG1	chr3	49482715	49489341	+	15	0.71
EBPL	chr13	49163600	49135193	-	85	0.25
EPB41L1	chr20	34249194	34263711	+	4	-0.24
FAM176A	chr2	75641931	75606745	-	263	-0.21
FAM62B	chr7	158245032	158235100	-	310	0.58
FER1L3	chr10	95145970	95138776	-	969	-0.29
HMGN1	chr21	39641278	39639625	-	35	-0.27
HMGN1	chr21	39642134	39638941	-	49	-0.21
KIF13A	chr6	17880268	17872364	-	62	0.56
MARK3	chr14	103027866	103036289	+	136	-0.37
MARK3	chr14	103027866	103039370	+	206	-0.49
MARK3	chr14	103027866	103039370	+	206	-0.24
MATR3	chr5	138642638	138671307	+	1190	0.37
MBNL1	chr3	153645760	153648251	+	78	0.30
NFYA	chr6	41154745	41159908	+	52	-0.72
NIN	chr14	50296826	50291209	-	59	-0.41
NISCH	chr3	52489239	52493692	+	11	0.29
PBRM1	chr3	52571023	52559802	-	86	0.24
PLEKHA1	chr10	124176447	124179643	+	58	-0.31
PLEKHA1	chr10	124176447	124179643	+	58	-0.26
PLEKHM2	chr1	15918815	15924626	+	3	-0.44
PLOD2	chr3	147279733	147277258	-	193	-0.32
PPFIBP1	chr12	27720627	27723838	+	57	0.26
RBM3	chrX	48318499	48319750	+	16	0.52
SEC31A	chr4	83984685	83969176	-	373	0.29
SEC31A	chr4	83984685	83969176	-	373	0.37
SEC31A	chr4	83982657	83969176	-	300	0.67
SLK	chr10	105757924	105768036	+	256	0.72
SPAG9	chr17	46409580	46407130	-	72	-0.36
STX2	chr12	129849132	129842039	-	38	0.42
TEAD1	chr11	12840372	12843022	+	77	-0.28
UBXN11	chr1	26505713	26500771	-	2	0.25
WARS	chr14	99912416	99897797	-	171	-0.35
ZC3H11A	chr1	202031397	202037378	+	1668	0.25
ZNF584	chr19	63611929	63618824	+	4	-0.23

splicing patterns were assessed in the parental MDA-231 cells (RBM47 positive) and 831BrM2 cells (RBM47 negative) to identify those RBM47-dependent alternative splicing events that may contribute to the more aggressive metastasis seen in the absence of RBM47.

Analysis of the RNAseq data identified MDM4 exon 6 (ex6) as a negatively regulated, RBM47-bound cassette exon ($\Delta I = -0.63$, 145 RBM47 CLIP tags in $BC \geq 2$ tags). Figure 5.26A shows the genomic locus of the MDM4 gene, with normalized RNAseq reads from Tet-On control (red) and RBM47-WT (green), and RBM47 CLIP tags in $BC \geq 2$ tags clusters (black). RBM47 binding is present in the 5' flanking intron of the alternatively spliced ex6, as well as in the 3'UTR and several additional non-alternatively spliced exons of MDM4. Zooming in on the region of the alternatively spliced exon (Figure 5.26B) clearly shows reduction in RNAseq signal corresponding to ex6 in RBM47 expressing cells. Semi-quantitative RT-PCR of the region spanning MDM4 ex5-7 shows significant RBM47-dependent exclusion of ex6 in dox-treated RBM47-WT#10 compared to Tet-On control ($\Delta I = -0.27$, $p = 4.4 \times 10^{-4}$). The level of MDM4 ex6 exclusion correlates with RBM47 expression in MDA-231 and 831BrM2 cells ($\Delta I = IR_{MDA} - IR_{BrM2} = -0.12$, $p = 1.4 \times 10^{-3}$, Figure 5.26C). RBM47-dependent splicing changes associated with intronic RBM47 binding were validated for alternative exons in MBNL1, MACF1 and KIFAP3 (Figure 5.27) with the anticipated changes in MDA-231 and 831BrM2 cells with respect to RBM47 expression. Several alternative exons showing robust RBM47 dependent changes in RBM47-WT#10 failed to show any splicing

FIGURE 5.26: RBM47 binding correlates with RBM47-dependent exon exclusion in MDM4. A: The genomic MDM4 locus showing RBM47 HITS-CLIP binding in BC2 \geq 2tags clusters in black and normalized RNAseq tracks from non-RBM47 expression Tet-On control (red) and RBM47-WT cells (green). RBM47 binding is evident in the 3'UTR of MDM4, and in the intronic region between constitutive ex5 and the alternatively spliced ex6. **B:** Close-up of the ex5-ex7 region of MDM4 showing clear decrease in RNAseq signal corresponding to ex6 in RBM47-WT cells. Shown below the gene are the cassette exons used for RNAseq alignment and IR calculation. Two included isoforms are present including full or partial ex6 (red) and the ex6 excluded isoform (blue). **C:** Independent verification of splicing carried out by semi-quantitative radioactive RT-PCR amplifying a region spanning ex5-ex7 in dox-treated Tet-On and RBM47-WT#10 (top left panel), and MDA-231 (RBM47 positive) and 831BrM2 (RBM47 negative, top right panel). In all splicing RT-PCRs the first three lanes consist of an equal-parts mixture of cDNA from all samples amplified at increasing cycle number to ensure amplification in the linear range, and the fourth lane contains a –RT control from the same mix. The middle cycle number was used to amplify cDNA from samples prepared in triplicate from each cell line. The presence of both exogenous and endogenous RBM47 correlates with preferential ex6 skipping. Average ΔI of RT-PCR reactions were calculated using ImageJ analysis of autoradiograms ((Schneider et al., 2012), lower panels), with p-value *** \leq 0.001, ** \leq 0.01 calculated by t-test.



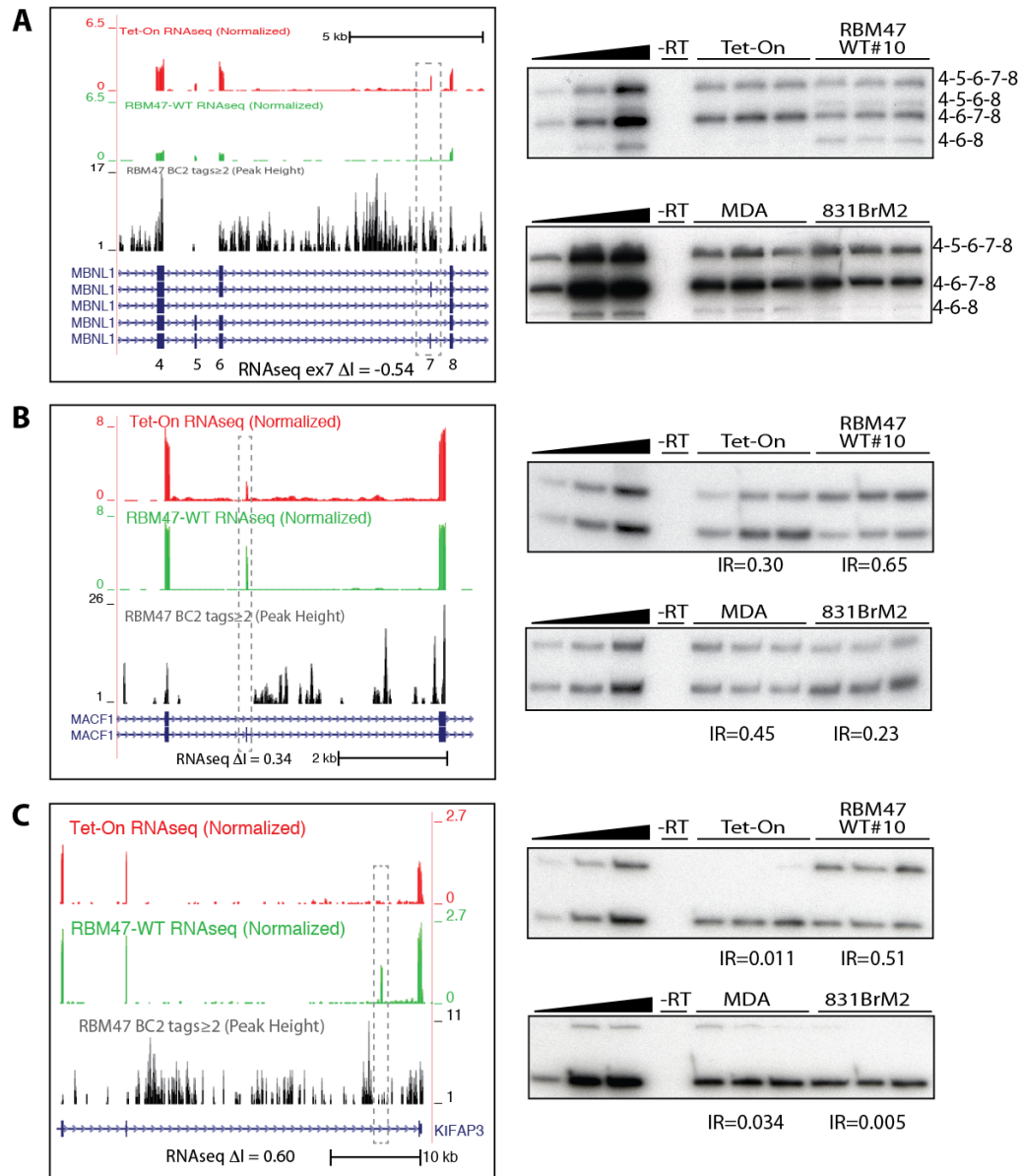


FIGURE 5.27: RBM47 expression correlates with RBM47-dependent exon inclusion and exclusion in bound targets. A: RBM47 binding in the region spanning ex4-ex8 of MBNL1 correlates with RBM47-dependent exclusion of ex7, an NMD exon, in both RBM47-WT#10 and MDA-231 cells. **B:** RBM47 promotes inclusion of ex91 of MACF1. RT-PCR $\Delta I_{\text{RBM47-WT\#10-TetOn}} = 0.35$, $p = 0.0003$; $\Delta I_{\text{MDA-831BrM2}} = 0.22$, $p = 0.00003$. **C:** RBM47 promotes the inclusion of an unannotated exon between ex19-20 in KIFAP3. Inclusion rate (IR) of RT-PCR reactions were calculated using ImageJ analysis of autoradiograms, p-value calculated by t-test where possible.

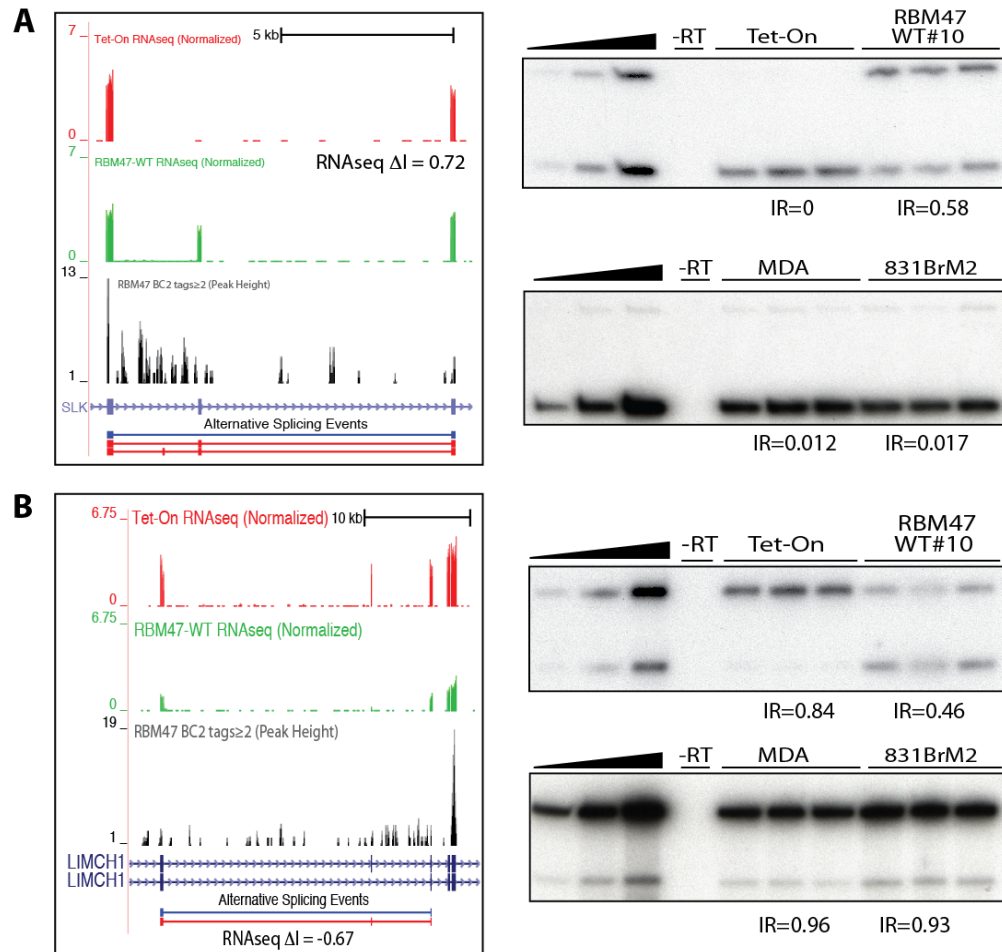


FIGURE 5.28: Exogenous RBM47 expression correlates with RBM47-dependent exon inclusion and exclusion not seen when comparing MDA-2321 and 831BrM2. A: RBM47 binding in the intron between ex12 and ex13 promotes ex13 inclusion in SLK. This exon is excluded in both MDA-231 and 831BrM2. **B:** LIMCH1 ex9 (based on variant 1) shows robust RBM47-dependent exclusion in RBM47-WT ($\Delta I_{\text{RBM47-WT\#10-TetOn}} = -0.53$, $p = 2.18 \times 10^{-5}$), however no change in splicing pattern was seen between MDA-231 and 831BrM2. Inclusion rate (IR) of RT-PCR reactions were calculated using ImageJ analysis of autoradiograms, p-value calculated by t-test where possible.

changes between MDA-231 and RBM47 (Figure 5.28). Together these results suggest a regulatory role for RBM47 in alternative splicing, with direct binding correlating with both exon inclusion and exclusion.

RBM47 alternative splicing RNA Binding Map

By coupling protein-dependent splicing changes to the binding sites derived from HITS-CLIP it is possible to map on a composite pre-mRNA patterns of binding associated with RBP-binding dependent exon inclusion and exclusion. To generate an RBM47 RNA binding map, the total number unique RBM47 CLIP tags within 1kb of each RBM47-bound alternatively spliced exon-intron junction ($|\Delta I| > 0.2$, $BC2 \geq 2$ tags cluster) were counted and assigned a weight to normalize for gene expression level, then assigned to the closest splice site. For each 50nt window, the weighted sum of CLIP tags was normalized by dividing by the total number of tags mapping onto the cassette exon and flanking introns and plotted as a function of distance from respective exon/intron junctions to generate normalized complexity map (Ince-Dunn et al., 2012, Licatalosi et al., 2008) (Figure 5.29).

Unlike the RNA maps generated for other alternative splicing regulators such as Nova (Licatalosi et al., 2008) and FOX2 (Yeo et al., 2009), the RBM47 normalized complexity map is strikingly symmetrical with respect to binding around both included and excluded intron. While some preferential binding to regions spanning the 250nt immediately downstream of the 3' FE and to the 3'FE itself is evident in those exons for

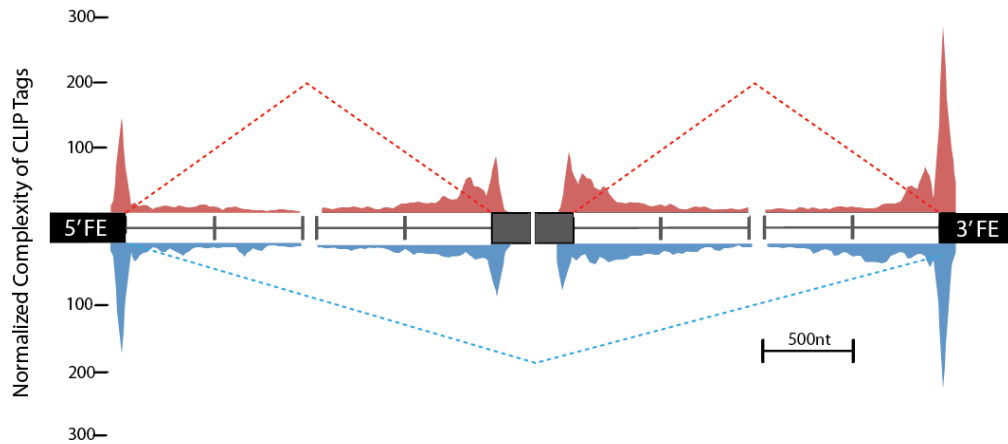


FIGURE 5.29: RBM47-RNA normalized complexity splicing map. All RBM47 unique CLIP tags mapping to the 944 identified alternative exons showing RBM47-dependent splicing changes and the presence of at least one $BC2 \geq 2$ tags cluster were placed onto a single composite pre-mRNA. Tags numbers were normalized both to the number and distribution of tags between transcripts, and to the number of transcripts showing binding at any given position to determine common potential regulatory sites. This map was generated twice, once for all included exons (red), and then for all excluded exons (blue). The normalized complexity binding map of RBM47 is strikingly symmetrical both in terms of binding in upstream and downstream regions relative to the cassette exons, and with respect to RBM47-dependent inclusion or exclusion.

which RBM47 binding promotes exon inclusion (red), there is no clear bias for upstream binding in the immediate vicinity of intron-exon boundary of included alternative exons (see Nova binding map (Licatalosi et al., 2008)). Peaks of binding are also observed in the 5'FE and at both of the alternative intron-exon boundaries, extending 50nt into the exonic region and greater than 500nt out from the junction in both the upstream and downstream directions. A tail of low-level binding can be seen extending into more distal intronic regions with a relatively constant normalized peak height. Surprisingly, the RBM47 binding map for exon skipping (blue) was remarkably similar to that of the RBM47 included exons, with near symmetrical binding at the exon-intron boundaries and flanking intronic regions and no bias towards binding in downstream regions with respect to the cassette exon.

The RBM47 binding seen in validated alternatively spliced transcripts occurs in multiple discrete clusters spanning the entire intron rather than occurring in clusters in close proximity to the splice site. Given the extended tails of binding seen in the RBM47 normalized complexity map, it was feasible that a bias towards upstream binding in excluded and downstream binding in included cassette exons could exist outside the 1kb boundary of the RNA map. In order to look for binding patterns across large intronic distances, RBM47 tags in $BC2 \geq 2$ tags clusters located in the 5'FE- 3'FE region were counted, and a percentage binding in each sub-region calculated. This percentage was then averaged across all RBM47-dependent included or excluded alternative events (Figure 5.30).

No clear bias towards upstream or downstream binding was evident in the RBM47 HITS-CLIP profile. Additionally, no bias towards binding one or both introns with respect to inclusion or exclusion was detected. It has been suggested that internal cassette exons are predominantly excluded when flanked by two long introns while the presence of a shorter neighboring intron can rescue exon inclusion (Fox-Walsh et al., 2005), however no bias was found in distribution of intron lengths or ratio of intron lengths around RBM47-regulated cassette exons (data not shown).

It remains a possibility that due to the single replicate RNAseq experiment and the relatively low threshold set for evidence of RBM47 binding noise is present in the current data set that obscures the RNA binding “rules” associated with RBM47 regulated alternative splicing. Increasing the stringency of identification of altered transcripts with replicate RNAseq libraries and increasing the threshold set for evidence of direct RBM47-binding, coupled with further RT-PCR validation should refine the list of regulated events to allow for further investigation into the rules governing RBM47-mediated inclusion or exclusion of alternative exons.

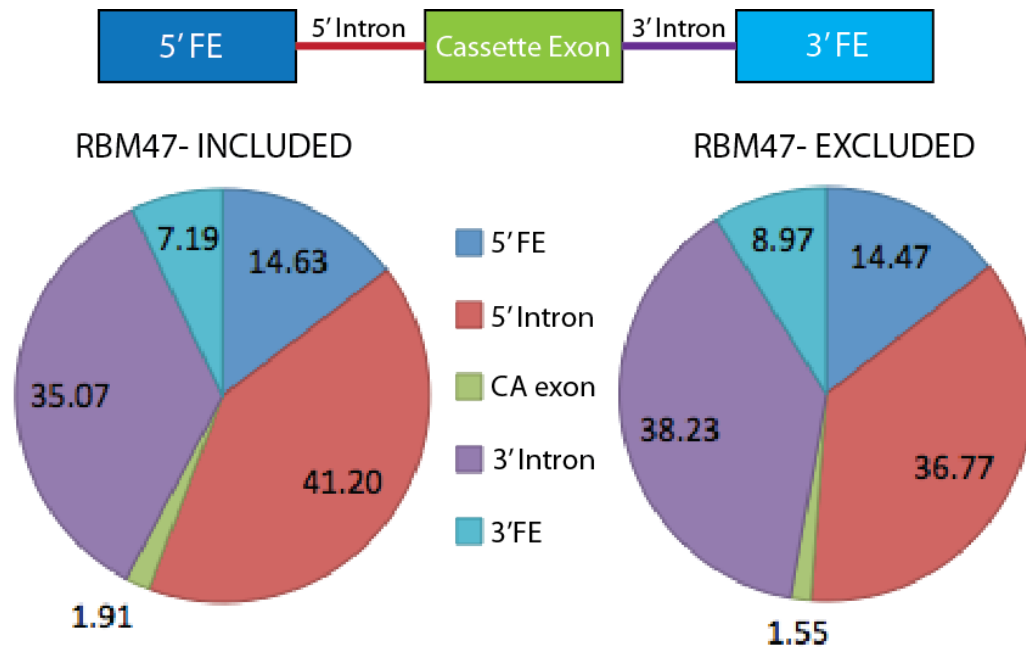


FIGURE 5.30: No extended positional bias is evident in RBM47 binding flanking regulated alternative exons. The percentage of RBM47 CLIP tags in BC2≥2tags clusters mapping to the 5' flanking exon, 5' intron, cassette exon, 3' flanking intron and 3' flanking exon were averaged across all RBM47-dependent inclusion and exclusion events. The average distribution of CLIP tags around the cassette exon is no different between included and excluded events.

DISCUSSION

In collaboration with the Massagué laboratory at MSKCC, we have been shown that RBM47 is a novel, potent breast cancer metastasis suppressor, and that decrease in RBM47 expression is clinically associated with poor metastasis-free survival to both brain and lung. Understanding the normal function of RBM47 is key to understanding how this anti-metastatic effect is mediated. In this study we have validated the RNA binding ability of RBM47 and, by mapping the entire network of RBM47-bound transcripts, we have determined that RBM47 is a multifunctional RBP with roles in both transcript stabilization and the regulation of alternative splicing.

Motif analysis of robust RBM47 CLIP clusters revealed a preference for polyU-rich sequences, consistent with a Class III AU-rich binding element (ARE). Class I AREs contain 1-3 copies of an AUUUA pentamer surrounded by U-rich regions, class II contain multiple overlapping copies of AUUUA (usually 5-8), while class III are more loosely defined in that they lack the AUUUA pentamer but instead contain predominantly U-rich sequence (Chen and Shyu, 1995). ELAVL1/HuR, by binding AREs in the 3'UTRs of GM-CSF and c-fos (Vakalopoulou et al., 1991) can promote increased half-life of both transcripts ((Fan and Steitz, 1998). By coupling HITS-CLIP binding maps with RNAseq analysis of RBM47-dependent steady state transcript level changes, we have identified a subset of targets in which robust RBM47 3'UTR binding correlates with both up regulation of steady state transcript levels and, in the case of DKK1, stabilization of

the transcript. The data presented here would suggest that RBM47, like ELAVL1 functions as a transcript stabilizing 3'UTR ARE-binding protein.

PolyU binding sites have been identified via HITS-CLIP (or its variations) for hnRNP-C (iCLIP (Konig et al., 2010)), ELAVL1 (PAR-CLIP, (Lebedeva et al., 2011, Mukherjee et al., 2011), HITS-CLIP Darnell lab unpublished data), nELAVL (Ince-Dunn et al., 2012), U2AF65 (Rogelj et al., 2012), and PTBP2 (Licatalosi et al., 2012). Such low complexity targeting is difficult to reconcile with the specificity of target selection shown for each RBP, indeed for many low complexity motif binding proteins the presence of a canonical motif is neither necessary or sufficient to predict binding (for example, binding of FUS to its consensus motif GGUG in both mouse and human brain (Lagier-Tourenne et al., 2012)). This would suggest other modulatory factors in target choice such as RNA accessibility and secondary structure (Hoell et al., 2011) or recruitment via protein-protein interactions. Identification of RBM47 interacting proteins may help to further elucidate the mechanism by which specific targets are selected for stabilization or alternative splicing.

In order to identify RBM47-bound transcripts whose RBM47-dependent regulation has the potential to influence metastatic progression, transcript levels were assessed both in the exogenous over-expression setting (RBM47-WT versus Tet-On control cells) and in the endogenous setting (MDA-231 versus 831BrM2). We have identified three RBM47-bound targets whose transcript levels positively correlate with

RBM47 expression: CTGF, ANKRD1 and DKK1. Given the stringent cut-offs placed on RBM47-dependent expression level changes as measured by RNAseq this is unlikely to be an exhaustive list of metastasis associated RBM47-regulated transcripts. Interestingly, for each of the RBM47 targets validated a decrease in the total AGO binding in 3'UTRs was observed in MDA-231 cells compared to 831BrM2, suggesting that RBM47 may mediate its stabilizing effect in part by displacing AGO from regulatory regions on common targets.

The complex interplay of RBPs in both agonistic and antagonistic modulation of miRNA regulation is becoming increasingly apparent. For example, PUF proteins have been shown to activate miRNA-mediated silencing of *hbl-1* by let-7 in *C.elegans* (Nolde et al., 2007) and to facilitate miR-221/222 silencing of p27 in human fibroblasts by binding to and altering local 3'UTR secondary structure (Kedde et al., 2010). As discussed in Chapter 3, the RRM domain containing ELAVL family members have been ascribed both agonistic, required for let-7 mediated c-Myc repression (Kim et al., 2009a), and antagonistic roles in AGO-mediated transcript regulation, binding to miR-16 to inhibit targeting of COX-2 (Young et al., 2012) and stress-mediated relief of CAT-1 mRNA repression by miR-122 (Bhattacharyya et al., 2006). For c-Myc and CAT-1 regulation, regulatory HuR 3'UTR binding sites are located in regions not overlapping targeting miRNA seed sites, consistent with PAR-CLIP data suggesting enrichment of miRNA seed sequences adjacent to but not within HuR binding sites (Lebedeva et al., 2011). *In vitro* studies have shown that HuR

oligomerization is required for the attenuation of miRISC target cleavage through displacement of the RISC complex from target mRNA (Kundu et al., 2012). Spreading of HuR binding along transcripts may help to explain more recent PAR- and HITS-CLIP data that have indicated a significant enrichment for overlapping HuR and AGO binding sites within 10nt of each other suggesting direct competition for binding sites ((Mukherjee et al., 2011), Darnell Lab unpublished data). The peak overlap of RBM47 and AGO in 831BrM2 cells does not share this discrete enrichment, rather RBM47 binding is distributed more evenly throughout AGO peaks suggesting a potentially less site specific co-regulation of targets.

The RRM domains of RBM47 bear high similarity (63.4%, (Rice et al., 2000)) to the single RRM domain containing protein dead end 1 (Dnd1), a U-rich element binding RBP predominantly expressed in primordial germ cells (Youngren et al., 2005). Endogenous Dnd1 expression in zebrafish germ cells and ectopic expression in human cells has been shown to broadly relieve miRNA repression through direct binding of target mRNA in regions proximal to or partially overlapping the targeting miRNA seed sequence (Kedde et al., 2007). Modulation of mRNA repression without changes in miRNA expression level has been observed in germ cells of *Arabidopsis thaliana* (miR-159, (Alonso-Peral et al., 2012)), zebra fish (miR-430, miR-221 / 222, miR-372, miR-1 and miR-206 (Kedde et al., 2007, Mishima et al., 2006)) and mouse oocytes where deletion of Dicer1 (Ma et al., 2010a) or Dgcr8 (Suh et al., 2010) fails to lead to de-repression of mRNA levels. Downregulation of DND1 has been

associated with increased target sensitivity to miR-21 in transformed keratinocytes (Bhandari et al., 2013) and in increased proliferation with reduced apoptosis in oral squamous cell carcinoma of the tongue (Liu et al., 2010), suggesting that the miRNA modulatory function of Dnd1 is not limited to germ cells and can be tumor suppressive in transformed somatic cells.

If RBM47 is able to modulate accessibility of AGO to common target 3'UTRs, loss of RBM47 in metastatic progression may allow for gain of function for multiple miRNAs by unmasking regulatory seed sites without requiring extensive alteration of miRNA expression patterns. Such transcript-wide increase in AGO binding in the absence of RBM47 can be inferred from the AGO binding maps of MDA-231 and 831BrM2, although this is highly correlative data. It would be interesting to assess direct AGO HITS-CLIP binding maps in the presence and absence of RBM47, either by RBM47-WT +/-dox induction or by RBM47 knockdown in MDA-231. Careful normalization of AGO binding with respect to target expression levels should enable assessment of RBM47-dependent disruption of AGO binding both on a genome-wide and RBM47-target specific level. Furthermore, given that AGO HITS-CLIP in 831BrM2 cells has already delimited the discrete AGO binding sites on targets such as DKK1, ANKRD1 and CTGF, the relative ability of individual seed-matched miRNA to repress target transcripts in the presence or absence of RBM47 can be assessed. Of note, additional miR-335/* mimic transfection into MDA-231 failed to decrease relative luciferase activity in full length

DKK1 3'UTR reporter assays (data not shown), consistent with a buffering of transcript level by endogenous RBM47 and suggesting that knockdown of RBM47 in MDA-231 cells may be required to unmask the regulatory potential of miR-335/* on the DKK1 transcript.

Defining high confidence RBP targets by HITS-CLIP has relied upon stringent filtering to discern robust and reproducible binding sites. Non-reproducible binding, or reproducible binding with low tag numbers, could stem from alignment errors, low-affinity transient binding captured by crosslinking that presumably does not represent the major function of the RBP, or in the case of Flag-RBM47 may reflect fortuitous binding events occurring due to non-stoichiometric levels of protein and target RNA. While requiring reproducibility between replicate HITS-CLIP libraries at any given binding site reduces the potential for background noise, applying increasingly stringent thresholds to RBM47 clusters lead to a dramatic loss in the number of retained intronic clusters and a relative enrichment in 3'UTR binding. A similar shift in proportional binding has been reported for nELAVL with increasing stringency placed on biological complexity of peaks, but importantly, intronic nELAVL binding has been shown demarcate functional sites of regulation with respect to alternative splicing (Ince-Dunn et al., 2012).

HITS-CLIP derived RNA binding maps associated with the inclusion or exclusion of alternatively spliced cassette exons have revealed a strong bias towards exon exclusion when binding occurs in the cassette

exon or upstream flanking intronic sequence, and exon inclusion with evidence of cassette exon-adjacent downstream intronic binding. Such patterns have been described for Nova (Ule et al., 2003, Ule et al., 2006, Licatalosi et al., 2008), FOX1/2 (Yeo et al., 2009), nELAVL (Ince-Dunn et al., 2012), Mbnl-1 (Wang et al., 2012) and Mbnl-2 (Charizanis et al., 2012) leading to the suggestion that this position dependent regulation may be a common feature of RBPs involved in alternative splicing regulation (Licatalosi and Darnell, 2010). Nova has been shown to promote exon skipping by inhibiting the action of U1 snRNP (Ule et al., 2006), therefore the discrete intronic binding seen in the immediate vicinity of alternatively spliced cassette exons and flanking constitutive exons may reflect the mechanism by which this and other RBPs regulate alternative splicing.

The binding map of RBM47 does not show clear up/downstream binding bias within the 500nt flanking alternatively splice exon-intron junctions and is instead strikingly symmetrical with respect to cassette exon inclusion or exclusion. The lack of a positional relationship between RNA binding and splicing outcome has previously been described in mouse embryonic stem cells for the well-characterized splicing factors SRSF1 and SRSF2, although for both of these proteins enrichment in binding was also seen on the constitutive exons consistent with binding to exonic splicing enhancers (Pandit et al., 2013). Relatively symmetrical normalized complexity maps have been generated from HITS-CLIP data for both PTBP1 (Xue et al., 2009) and PTBP2 (Licatalosi et al., 2012),

however these maps demonstrate some asymmetric binding enrichment at intron-exon boundaries and lack the long tails of intronic binding seen with RBM47.

Most similar to the RBM47-intronic binding patterns described here are those associated with alternative splicing of FUS and TDP-43. For both of these proteins multiple reproducible intronic peaks of low total tag number (in many cases below the 10 tag threshold at which ~90% of RBM47 intronic signal is in lost) are correlated with respective protein-dependent regulation of alternative splicing ((Tollervey et al., 2011, Rogelj et al., 2012, Lagier-Tourenne et al., 2012). It has been suggested that binding to long introns in regions greater than 2kb from the exon-intron boundary is a salient feature of both TDP-43 and FUS regulation of pre-mRNA stability and splicing (Lagier-Tourenne et al., 2012, Polymenidou et al., 2011), however alternative splicing of cassette exons in both human and *Drosophila* is generally associated with long intron length (Fox-Walsh et al., 2005). The median length of RBM47-bound introns was found to be significantly longer than the total population of both constitutive and cassette exon-associated introns, with bias towards binding in the distal regions (data not shown). However, the same significant increase in bound intron length was also found when analyzing AGO intronic binding in 831BrM2 cells, suggesting that the preference for binding to long introns is not a unique feature of RBM47.

RBM47 intronic binding is associated with extensive patterns of altered alternative splicing, with several of the candidate RBM47-regulated transcripts validated here also showing loss of RBM47-specific alternative splicing patterns in 831BrM2 compared to MDA-231. This would suggest that loss of RBM47 pre-mRNA splicing regulation might contribute towards enhanced metastatic potential in 831BrM2 cells. Shapiro et al. (2011) have reported extensive alternative splicing associated with a breast cancer cell line model of EMT, the majority of which they attribute to down regulation of PTB and ESRP1/2, and up regulation of RBFOX2 through comparison with previously published CLIP data sets to enrichment of binding motifs within a 250nt flanking region of regulated splice sites. RBM47 is also downregulated in this system, and based on the HITS-CLIP and RNAseq data presented here, has the potential to directly regulate approximately 25% of the reported EMT-associated splicing changes. The polyU RBM47 enriched-motif derived from HITS-CLIP clusters was not among the significantly enriched 5-mers in the sequences adjacent to EMT-associated splices, however this is not unexpected given the relatively deep intronic RBM47-binding patterns discussed above. Together this data strongly suggests that RBM47 is a novel regulator in the EMT-associated splicing program, whose downregulation may promote a more invasive phenotype in 831BrM2 cells.

The RBM47-regulated transcript MDM4 encodes a p53 binding protein with significant homology to MDM2, a potent negative inhibitor

of p53 (Shvarts et al., 1996). Multiple alternatively spliced MDM4 transcript variants have been described (reviewed in (Mancini et al., 2009)), with exclusion of the RBM47-regulated exon 6 leading to a truncated but functional variant termed MDM4-S. MDM4-S associates with p53 with a 10-fold higher affinity than full-length MDM4 leading to enhanced suppression of p53 transcriptional activity, making MDM4-S a potent oncogene (Rallapalli et al., 1999, Rallapalli et al., 2003), however an opposing role for MDM4-mediated p53 stabilization under conditions of cellular stress has also been described (Di Conza et al., 2012). How MDM4-S and MDM4 interact with the mutant p53 expressed in MDA-231 is unknown. It has been shown that decreased Mdm4 expression in mice can improve overall survival in a WT p53 Rb^{+/}- background, but not in a hypomorphic p53 mutant Rb^{+/}- background where the mutant p53 has partial capacity to induce apoptosis but is an inefficient transcriptional activator (Fang et al., 2013b). Full length MDM4 has been reported to interact with multiple other proteins, including MDM2, p21 and CK1 α , using domains deleted from the truncated MDM4-S. These domains also contain multiple targets for post-translational modification (Mancini et al., 2009), therefore any potential pro-metastatic benefit of decreasing MDM4-S in 831BrM2 may be independent of any p53 regulation.

RBM47 was found to bind to and regulate the alternative splicing of several other known RBPs that themselves have alternative splicing regulatory functions, including MBLN1 and PTBP2 (data not shown). MBNL1 is abundantly expressed in skeletal muscle, brain and heart, and

has been implicated in the genetic disease myotonic dystrophy (DM1) (Miller et al., 2000). RBM47 intronic binding is associated with inclusion of exon 7. Exon 7 is required for MBNL1 dimerization and enhances but is not required for binding to CUG repeats (Tran et al., 2011), however inclusion or exclusion does not alter MBNL1 splicing activity (Lin et al., 2006). Interestingly, inclusion of exon 7 represents a fetal isoform of MBNL1, with failure of postnatal downregulation of exon 7 inclusion seen in mouse models of DM1 and human DM1 skeletal muscle tissue (Lin et al., 2006).

That RBM47 is able to regulate other splicing factors may in part explain the presence of RBM47-dependent alternatively spliced transcripts that do not show evidence of direct RBM47 binding by HITS-CLIP. It will be interesting to investigate whether modulation of individual transcript RBM47-dependent alternative splicing is able to enhance metastatic ability. Given the relatively dispersed RBM47 binding observed in regulated introns, using oligonucleotides targeted to block RBM47 binding in MDA-231 cells may be difficult, however steric-blocking oligonucleotides may be of use to prevent RBM47-enhanced cassette exon inclusion (reviewed in (Kole et al., 2012)).

The tissue-specific expression pattern of endogenous RBM47 is currently unknown. BioGPS (Wu et al., 2009a) reports high expression in thyroid, prostate, small intestine and colon in addition to several immune cell types, while kidney has the highest expression (RPKM) seen in the

RNAseq Atlas (Krupp et al., 2012). Interpretation of the cytoplasmic subcellular localization of RBM47 reported in the Human Protein Atlas (Uhlen et al., 2010) is limited by the use of a single antibody known to have non-specific cross reactivity. The extensive intronic binding evident in the HITS-CLIP binding map and the role of RBM47 in regulating alternative splicing of a subset of targets would indicate that RBM47 is at least in part nuclear. Development of an antibody with high specificity and affinity for RBM47 will be crucial to establishing the expression patterns of RBM47 in normal and disease tissues, to determine endogenous subcellular localization and for endogenous HITS-CLIP.

CHAPTER 6: GENERAL DISCUSSION

Summary

HITS-CLIP has been used to study RNA regulation in a variety of systems, including neurons, T cells and stem cells, but prior to this study HITS-CLIP had not been applied to study RNA regulation in human cancer. Here we have described the first study of differential RNA regulation with respect to the acquisition of the aggressively metastatic phenotype, using HITS-CLIP as a genome-wide approach to identify AGO-regulated targets. In this study, we have assessed the role of RNA binding proteins in metastasis suppression using an elegant human xenograft based system developed by the Massagué laboratory that accurately models breast tumor dissemination and metastatic colonization of distal tissues (Bos et al., 2009, Kang et al., 2003, Minn et al., 2005a, Minn et al., 2005b). This system has previously been used to identify anti-metastatic miRNAs whose expression levels correlate with human metastatic relapse (Png et al., 2011, Png et al., 2012, Tavazoie et al., 2008), however the identity of directly regulated targets of these miRNAs is difficult to infer from correlative changes in expression levels and reporter assays.

Using AGO HITS-CLIP in MDA-231 breast cancer cells and its highly aggressive metastatic sub-lines with tissue tropism to lung, bone and brain, we have generated transcriptome wide maps of AGO-mRNA interactions which delineate sites of regulatory action of AGO on target

mRNA in a genome wide manner. In so doing, a valuable resource has been generated that allows for narrowing of the sequence space in which to search for regulatory miRNA seed sites on any given AGO-regulated transcript in MDA-231 cells. Additionally by comparing the binding maps generated in these four cell lines, we have identified sites of differential AGO-mRNA binding that correlate with changes in expression levels of target mRNAs with known roles in metastasis. Furthermore, through analysis of the simultaneously derived AGO-miRNA libraries from HITS-CLIP, we have identified several miRNAs with whose CLIP profiles are altered with respect to metastatic potential. Decrease in the AGO-miRNA counts of the miR-99a/let-7c/125b-2 cluster members were found to be specific to lung metastatic cells, while divergent decrease (lung) and increase (brain and bone) in the representation of miR-142-3p suggests further potential for miRNA regulation with respect to tissue tropism.

Unexpectedly, the miRNAs previously identified as endogenous suppressors of metastasis were weakly represented in the AGO HITS-CLIP profile of MDA-231 cells, indicating a previously unappreciated lower limit of sensitivity of the AGO HITS-CLIP approach. Therefore to identify the full repertoire of regulated transcripts for one such anti-metastatic miRNA, differential AGO HITS-CLIP was carried out in lung metastatic cells stably expressing pri-miR-335, expression that is sufficient to significantly inhibit lung colonization, or a non-targeting control shRNA. As a result, we have uncovered a previously unappreciated regulatory role for miR-335*, the presumed passenger strand of pre-miR-

335, expanding the potential for anti-metastatic target recognition and potential therapeutic modulation of metastasis. Additionally, we have identified several hundred transcripts miR-335/* seed containing *de novo* AGO binding sites, with the validation of several previously unknown miR-335/* targets providing proof of principle that *de novo* AGO-mRNA binding is an indicator of gain of miR-335/* regulation.

By applying HITS-CLIP prospectively to a previously uncharacterized protein whose downregulation in primary breast tumor correlates with poor metastasis-free survival to lung and brain, RBM47, we have demonstrated that is a multifunctional RNA binding protein with roles in both transcript stability and alternative splicing. Through analysis of robust and reproducible RBM47 RNA binding sites we have shown a preference for RBM47 binding to AU-rich sequences, with the most robust binding occurring in the 3'UTRs of target mRNA. In a sub-set of these targets, including DKK1, ANKRD1 and CTGF, we have demonstrated that RBM47 binding is positively correlated with increase in target mRNA steady state levels, which in the case of DKK1 was shown to be due to enhanced transcript stability. Furthermore we have established a role for RBM47 as a regulator of alternative splicing, with binding of RBM47 in intronic regions flanking RBM47-dependent alternative splice cassette exons promoting both exon inclusion and exclusion.

Comparison of AGO and RBM47 HITS-CLIP binding maps reveal convergence on regulatory nodes of the Wnt pathway

While each dataset generated by HITS-CLIP is informative with respect to the individual RBP and targets it regulates, by undertaking multiple experiments within the same relatively homogenous population of cells it is possible to overlay binding maps to identify regions of potential agonistic or antagonistic regulation. By comparing AGO and RBM47 binding maps in 831BrM2 cells the 3'UTR of DKK1 was identified as a common target. DKK1 is a member of the DKK family of secreted Wnt inhibitors that binds to and sequesters the Wnt co-receptors LRP5/6, and can interact with Kremen1/2 to block Wnt signaling (Niehrs, 2006). Two other nodes of the canonical Wnt pathway were identified as either targets of RBM47 splicing or as miR-335 responsive genes (Tavazoie et al., 2008, Oskarsson et al., 2011) (Figure 6.1).

Wnt signaling is required for mouse mammary stem cell renewal (Alonso-Peral et al., 2012) and maintenance of the cancer stem cell phenotype in murine and human squamous cell carcinomas (Malanchi et al., 2008). Interestingly, Wnt signaling is required for metastatic outgrowth of breast cancer cells in the lung via periostin recruitment of Wnt ligands from the stroma in the murine MMTV-PyMT breast cancer model (Malanchi et al., 2012). In human epithelial stem cells (hESC) Wnt signaling has been proposed to promote survival and differentiation (Blauwkamp et al., 2012), however other reports indicate that Wnt signaling is required for self-renewal of hESCs and inhibition of

differentiation (Singh et al., 2012). The pro- or anti-metastatic role of Wnt signaling and DKK1 expression is both tumor type and distal-organ specific (Menezes et al., 2012). As a negative regulator of bone formation (Pinzone et al., 2009), DKK1 expression has been found to enhance bone metastasis of multiple tumor types including prostate (Hall et al., 2005) and breast (Bu et al., 2008). Conversely, DKK1 secreted from human mesenchymal stem cells can inhibit MCF-7 breast cancer cell growth (Qiao et al., 2008), while down regulation of DKK1 enhances the motility, foci formation and soft agar colonization of nontumorigenic-epithelial breast cells, and leads to partial EMT (Mitra et al., 2010).

We have shown that DKK1 expression is decreased in lung and brain metastatic sub-lines of MDA-231, an expression level change that was not previously reported as part of respective metastasis signatures due to stringent >3-fold cut-offs of expression level change. DKK1 expression is known to be regulated at the transcriptional level by canonical Wnt signaling (Gonzalez-Sancho et al., 2005), vitamin D3 (Aguilera et al., 2007), and by epigenetic silencing through CpG island hypermethylation and histone deacetylation in multiple human cancers including colorectal (Aguilera et al., 2006) and breast (Suzuki et al., 2008). Post-transcriptionally, DKK1 has been reported to be negatively regulated by a family of related miRNAs (miR-93.hd/291-3p/294/295/302/372/373/520; (Zovoilis et al., 2009, Zhou et al., 2012)), miR-335 (Zhang et al., 2011) and miR-29a (Kapinas et al., 2010), with AGO binding in MDA-231 and derivatives supporting a role for the first two

regulatory events in this system. The data presented here indicates that RBM47 is a novel, direct positive post-transcriptional regulator of DKK1. Preliminary data from our collaborators at MSKCC indicates an RBM47 dose-dependent decrease in the expression of the endogenous Wnt target AXIN2 after Wnt3a stimulation (data not shown). Taken together, these data suggest downregulation (4175LM2) or loss of RBM47 (831BrM2) may lead to decreased DKK1 expression and increased activation of the Wnt pathway in metastatic derivatives.

We also identified DKK1 as a robust target for AGO regulation, with two of the AGO-binding sites shown to be pre-miR-335 dependent. No impact on DKK1 transcript level or 3'UTR luciferase reporter activity was evident upon miR-335/* mimic transfection suggesting that DKK1 levels may already be maximally downregulated in these cells, that our assay was not sensitive enough to detect subtle changes or that this 3'UTR may function as a neutral target or competitive endogenous miRNA sponge. Pre-miR-335 is among a cohort of anti-tumorigenic miRNAs reported to down regulate the Wnt pathway in HEK293 cells, although the directly regulated targets in this system were not elucidated (Anton et al., 2011). Additionally, the miR-335 responsive gene TNC (Tavazoie et al., 2008) has been shown in 4175LM2 cells to enhance Wnt signaling through promotion of LGR5 expression, and this Wnt signaling is required for lung metastasis outgrowth (Oskarsson et al., 2011). LGR5 forms a complex with LRP6 and Fzd5 that is rapidly internalized and degraded upon co-stimulation with R-spondin1 and Wnt3a (Carmon et al., 2012). TNC is

itself a Wnt target gene, with activated β -catenin upregulating TNC in embryonic mouse lung through direct binding to the TNC promoter (Cohen et al., 2009). Reduction of DKK1 in 4175LM2 cells may result in the upregulation of TNC in these cells via enhanced Wnt signaling, however the mechanism by which miR-335/* re-expression mediates TNC downregulation is unknown as analysis of AGO HITS-CLIP data indicates that TNC is not directly targeted by miR-335/*.

Additional links to Wnt signaling were found when analyzing the HITS-CLIP data associated with RBM47-bound alternatively spliced transcripts. The RBM47 splicing target microtubule actin cross-linking factor 1 (MACF1) is required for Wnt-mediated induction of TCF/LEF induction of Wnt target genes in Rat-1 embryonic fibroblast cells through proposed interactions with LRP6, APC, B-catenin and GSK3 β (Chen et al., 2006). In hair follicle stem cells Wnt signaling has been shown to inhibit GSK3 β phosphorylation of MACF1, promoting MACF1 association with microtubules and enhancing cell polarization and motility during wound repair (Wu et al., 2011b). RBM47 regulates both the inclusion of the alternatively spliced exon shown in Figure 5.27B, located between the EF-hand and Gas-2 related (GAR) domains of MACF1, and an additional exon between the GAR domain and the GSR-repeat domain (data not shown). The functional consequence of these in-frame exon inclusions on MACF1 activity is unknown. Given that RBM47 is known to be downregulated in EMT of breast cells (Shapiro et al., 2011) it is interesting to speculate that loss of RBM47 regulation of these exons might lead to

enhanced microtubule association and motility in metastasis, particularly when coupled with RBM47-enhanced DKK1 inhibition of Wnt signaling-dependent MACF1 phosphorylation. Taken together, the data presented here suggests that RNA binding proteins influence multiple nodes of the Wnt pathway. Concomitant loss of RBM47 and pre-miR-335 promote enhanced Wnt signaling, a previously unrecognized event in breast cancer metastasis (Figure 6.1).

Future Directions

In the work presented here, a cell culture model system was selected to interrogate the role of RNA binding proteins in metastasis. As a first step in the analysis of RNA regulation in metastatic versus primary cancers, the approach was chosen given the extensive previous molecular characterization of the xenograft and culture system, its demonstrated recapitulation of human breast cancer progression, and technical necessity due to the lack of RBM47 antibodies. However, multiple studies investigating a diverse array of RBPs have shown that HITS-CLIP can be successfully applied to whole tissues in order to identify *in vivo* sites of RNA regulation, with particular focus on neurological disorders. Preliminary studies obtained in the course of this thesis work have shown that it is also possible to apply HITS-CLIP technology to snap-frozen, human brain and to surgically resected tumors, specifically in the mapping of ectopic Nova on endogenous transcripts in neurons or in primary serous ovarian tumors, respectively (Christina Marney, Darnell Lab unpublished data). The ubiquitous expression of AGO and the

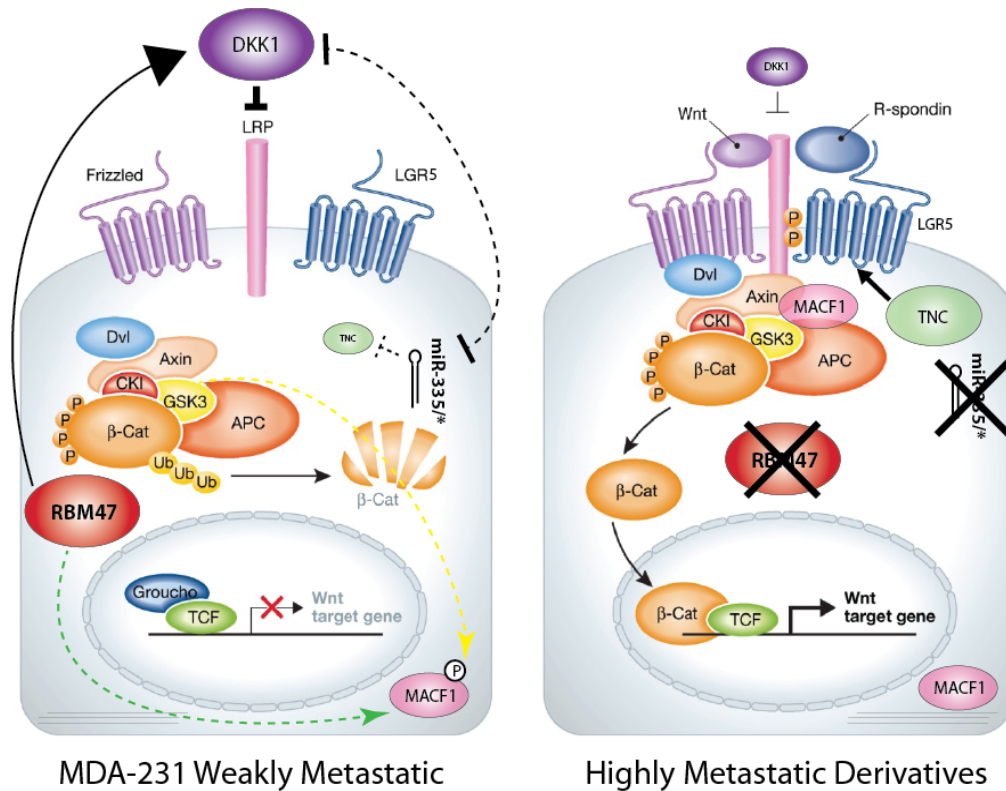


FIGURE 6.1: Multiple nodes of the DKK1/Wnt pathway are regulated by RBM47 and AGO-miR-335/* with respect to metastatic potential. In the parental MDA-231 cells that show a relatively poor ability to metastasize, RBM47 expression leads to stabilization and up regulation of DKK1 transcript, DKK1 binds to LRP5/6 leading to inhibition of the Wnt signaling pathway. This allows GSK3 β to phosphorylate the RBM47 splicing target MACF1 preventing it from associating with microtubules. Additionally, higher levels of pre-miR-335 leads to indirect repression of TNC expression, while DKK1 3'UTR is able to bind miR-335/* potentially without negative regulation of the DKK1 transcript. In the highly aggressively metastatic lung and brain metastatic sub-lines, both RBM47 and pre-miR-335 levels are reduced, leading to destabilization of the DKK1 transcripts and derepression of Wnt signaling. Loss of pre-miR-335 allows for TNC-dependent LGR5 expression leading to further potential enhancement of Wnt signaling, while loss of GSK3 β phosphorylation of MACF1 allows for association with microtubules that in hair follicle stem cells allows for cell polarization and directional movement. MACF1 may also promote translocation of Axin to the cell membrane. Figure adapted from (Schuijers and Clevers, 2012)

heterogeneous nature of primary tumors might make interpretation of tumor-derived AGO HITS-CLIP binding maps challenging. With careful selection of high density tumor-cell containing biological samples and paired analysis of normal adjacent tissues, and ongoing improvements in the sensitivity of the HITS-CLIP protocol, future studies of AGO and RBM47 with respect to metastasis progression may be possible in primary tumors and distal metastases.

REFERENCES

Aguilera O, Fraga MF, Ballestar E et al. Epigenetic inactivation of the Wnt antagonist DICKKOPF-1 (DKK-1) gene in human colorectal cancer. *Oncogene*. 2006;25:4116-21.

Aguilera O, Pena C, Garcia JM et al. The Wnt antagonist DICKKOPF-1 gene is induced by 1 α ,25-dihydroxyvitamin D3 associated to the differentiation of human colon cancer cells. *Carcinogenesis*. 2007;28:1877-84.

Alonso-Peral MM, Sun C, Millar AA. MicroRNA159 can act as a switch or tuning microRNA independently of its abundance in Arabidopsis. *PLoS One*. 2012;7:e34751.

Alt FW, Bothwell AL, Knapp M et al. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3' ends. *Cell*. 1980;20:293-301.

American Cancer Society. Breast Cancer Facts & Figures 2011-2012. Atlanta: American Cancer Society, Inc. 2012

Anczukow O, Buisson M, Leone M et al. BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. *Clin Cancer Res*. 2012a;18:4903-9.

Anczukow O, Rosenberg AZ, Akerman M et al. The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation. *Nat Struct Mol Biol*. 2012b;19:220-8.

Anczukow O, Ware MD, Buisson M et al. Does the nonsense-mediated mRNA decay mechanism prevent the synthesis of truncated BRCA1, CHK2, and p53 proteins? *Hum Mutat*. 2008;29:65-73.

Anton R, Chatterjee SS, Simundza J, Cowin P, Dasgupta R. A systematic screen for micro-RNAs regulating the canonical Wnt pathway. *PLoS One*. 2011;6:e26257.

Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature*. 2008;455:64-71.

Baffa R, Fassan M, Volinia S et al. MicroRNA expression profiling of human metastatic cancers identifies cancer gene targets. *J Pathol.* 2009;219:214-21.

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol.* 1994;2:28-36.

Bakheet T, Frevel M, Williams BR, Greer W, Khabar KS. ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res.* 2001;29:246-54.

Ball NS, King PH. Neuron-specific hel-N1 and HuD as novel molecular markers of neuroblastoma: a correlation of HuD messenger RNA levels with favorable prognostic features. *Clin Cancer Res.* 1997;3:1859-65.

Ballarino M, Jobert L, Dembele D, de la Grange P, Auboeuf D, Tora L. TAF15 is important for cellular proliferation and regulates the expression of a subset of cell cycle genes through miRNAs. *Oncogene.* 2012

Baltz AG, Munschauer M, Schwanhaussner B et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell.* 2012;46:674-90.

Bannai H, Fukatsu K, Mizutani A et al. An RNA-interacting protein, SYNCRIP (heterogeneous nuclear ribonuclear protein Q1/NSAP1) is a component of mRNA granule transported with inositol 1,4,5-trisphosphate receptor type 1 mRNA in neuronal dendrites. *J Biol Chem.* 2004;279:53427-34.

Barabino SM, Keller W. Last but not least: regulated poly(A) tail formation. *Cell.* 1999;99:9-11.

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* 2004;116:281-97.

Bertolotti A, Lutz Y, Heard DJ, Chambon P, Tora L. hTAF(II)68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. *EMBO J.* 1996;15:5022-31.

Bertolotti A, Melot T, Acker J, Vigneron M, Delattre O, Tora L. EWS, but not EWS-FLI-1, is associated with both TFIID and RNA polymerase II: interactions between two members of the TET family, EWS and hTAFII68, and subunits of TFIID and RNA polymerase II complexes. *Mol Cell Biol.* 1998;18:1489-97.

Bhandari A, Gordon W, Dizon D et al. The Grainyhead transcription factor Grhl3/Get1 suppresses miR-21 expression and tumorigenesis in skin: modulation of the miR-21 target MSH2 by RNA-binding protein DND1. *Oncogene.* 2013;32:1497-507.

Bhattacharyya SN, Habermacher R, Martine U, Closs EI, Filipowicz W. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell.* 2006;125:1111-24.

Blanc V, Navaratnam N, Henderson JO et al. Identification of GRY-RBP as an apolipoprotein B RNA-binding protein that interacts with both apobec-1 and apobec-1 complementation factor to modulate C to U editing. *J Biol Chem.* 2001;276:10272-83.

Blauwkamp TA, Nigam S, Ardehali R, Weissman IL, Nusse R. Endogenous Wnt signalling in human embryonic stem cells generates an equilibrium of distinct lineage-specified progenitors. *Nat Commun.* 2012;3:1070.

Blaxall BC, Dwyer-Nield LD, Bauer AK, Bohlmeier TJ, Malkinson AM, Port JD. Differential expression and localization of the mRNA binding proteins, AU-rich element mRNA binding protein (AUF1) and Hu antigen R (HuR), in neoplastic lung tissue. *Mol Carcinog.* 2000;28:76-83.

Bonnal S, Martinez C, Forch P, Bachi A, Wilm M, Valcarcel J. RBM5/Luca-15/H37 regulates Fas alternative splice site pairing after exon definition. *Mol Cell.* 2008;32:81-95.

Bos PD, Nguyen DX, Massague J. Modeling metastasis in the mouse. *Curr Opin Pharmacol.* 2010;10:571-7.

Bos PD, Zhang XH, Nadal C et al. Genes that mediate breast cancer metastasis to the brain. *Nature.* 2009;459:1005-9.

Boutz PL, Stoilov P, Li Q et al. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev.* 2007;21:1636-52.

Braeutigam C, Rago L, Rolke A, Waldmeier L, Christofori G, Winter J. The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. *Oncogene.* 2013

Bu G, Lu W, Liu CC et al. Breast cancer-derived Dickkopf1 inhibits osteoblast differentiation and osteoprotegerin expression: implication for breast cancer osteolytic bone metastases. *Int J Cancer.* 2008;123:1034-42.

Budde-Steffen C, Anderson NE, Rosenblum MK, Posner JB. Expression of an antigen in small cell lung carcinoma lines detected by antibodies from patients with paraneoplastic dorsal root ganglionopathy. *Cancer Res.* 1988;48:430-4.

Busch A, Hertel KJ. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA.* 2012;3:1-12.

Cailleau R, Olive M, Cruciger QV. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro.* 1978;14:911-5.

Calin GA, Ferracin M, Cimmino A et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med.* 2005;353:1793-801.

Calin GA, Sevignani C, Dumitru CD et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc Natl Acad Sci U S A.* 2004;101:2999-3004.

Cameron MD, Schmidt EE, Kerkvliet N et al. Temporal progression of metastasis in lung: cell survival, dormancy, and location dependence of metastatic inefficiency. *Cancer Res.* 2000;60:2541-6.

Carmon KS, Lin Q, Gong X, Thomas A, Liu Q. LGR5 interacts and cointernalizes with Wnt receptors to modulate Wnt/beta-catenin signaling. *Mol Cell Biol.* 2012;32:2054-64.

Castello A, Fischer B, Eichelbaum K et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell.* 2012;149:1393-406.

Chang S, Wang RH, Akagi K et al. Tumor suppressor BRCA1 epigenetically controls oncogenic microRNA-155. *Nat Med.* 2011;17:1275-82.

Chang TC, Yu D, Lee YS et al. Widespread microRNA repression by Myc contributes to tumorigenesis. *Nat Genet.* 2008;40:43-50.

Charizanis K, Lee KY, Batra R et al. Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy. *Neuron.* 2012;75:437-50.

Chen CY, Shyu AB. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci.* 1995;20:465-70.

Chen HH, Chang JG, Lu RM, Peng TY, Tarn WY. The RNA binding protein hnRNP Q modulates the utilization of exon 7 in the survival motor neuron 2 (SMN2) gene. *Mol Cell Biol.* 2008;28:6929-38.

Chen HJ, Lin CM, Lin CS, Perez-Olle R, Leung CL, Liem RK. The role of microtubule actin cross-linking factor 1 (MACF1) in the Wnt signaling pathway. *Genes Dev.* 2006;20:1933-45.

Chen L, Madura K. Centrin/Cdc31 is a novel regulator of protein degradation. *Mol Cell Biol.* 2008;28:1829-40.

Chen X, Yan Q, Li S et al. Expression of the tumor suppressor miR-206 is associated with cellular proliferative inhibition and impairs invasion in ERalpha-positive endometrioid adenocarcinoma. *Cancer Lett.* 2012;314:41-53.

Cheng C, Sharp PA. Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. *Mol Cell Biol.* 2006;26:362-70.

Cheng J, Zhou T, Liu C et al. Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science.* 1994;263:1759-62.

Chi SW, Hannon GJ, Darnell RB. An alternative mode of microRNA target recognition. *Nat Struct Mol Biol.* 2012;19:321-7.

Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature.* 2009;460:479-86.

Chiappetta G, Avantiaggiato V, Visconti R et al. High level expression of the HMGI (Y) gene during embryonic development. *Oncogene*. 1996;13:2439-46.

Choudhry H, Catto JW. Epigenetic regulation of microRNA expression in cancer. *Methods Mol Biol*. 2011;676:165-84.

Cohen ED, Ihida-Stansbury K, Lu MM, Panettieri RA, Jones PL, Morrissey EE. Wnt signaling regulates smooth muscle precursor development in the mouse lung via a tenascin C/PDGFR pathway. *J Clin Invest*. 2009;119:2538-49.

Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell*. 2009;136:777-93.
Cox JA, Tirone F, Durussel I et al. Calcium and magnesium binding to human centrin 3 and interaction with target peptides. *Biochemistry*. 2005;44:840-50.

Cui W, Zhao Y, Shan C et al. HBXIP upregulates CD46, CD55 and CD59 through ERK1/2/NF-kappaB signaling to protect breast cancer cells from complement attack. *FEBS Lett*. 2012;586:766-71.

D'Angelo D, Palmieri D, Mussnich P et al. Altered microRNA expression profile in human pituitary GH adenomas: down-regulation of miRNA targeting HMGA1, HMGA2, and E2F1. *J Clin Endocrinol Metab*. 2012;97:E1128-38.

Dalmau J, Furneaux HM, Cordon-Cardo C, Posner JB. The expression of the Hu (paraneoplastic encephalomyelitis/sensory neuronopathy) antigen in human normal and tumor tissues. *Am J Pathol*. 1992;141:881-6.

Dalmau J, Furneaux HM, Gralla RJ, Kris MG, Posner JB. Detection of the anti-Hu antibody in the serum of patients with small cell lung cancer--a quantitative western blot analysis. *Ann Neurol*. 1990;27:544-52.

Dalmau J, Graus F, Cheung NK et al. Major histocompatibility proteins, anti-Hu antibodies, and paraneoplastic encephalomyelitis in neuroblastoma and small cell lung cancer. *Cancer*. 1995;75:99-109.

Danckwardt S, Hentze MW, Kulozik AE. 3' end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J*. 2008;27:482-98.

Dantas TJ, Wang Y, Lalor P, Dockery P, Morrison CG. Defective nucleotide excision repair with normal centrosome structures and functions in the absence of all vertebrate centrins. *J Cell Biol.* 2011;193:307-18.

Darnell JC, Van Driesche SJ, Zhang C et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell.* 2011;146:247-61.

Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA.* 2010;1:266-86.

Darnell RB, Furneaux HM, Posner JB. Antiserum from a patient with cerebellar degeneration identifies a novel protein in Purkinje cells, cortical neurons, and neuroectodermal tumors. *J Neurosci.* 1991;11:1224-30.

Das S, Anczukow O, Akerman M, Krainer AR. Oncogenic splicing factor SRSF1 is a critical transcriptional target of MYC. *Cell Rep.* 2012;1:110-7.

David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.* 2010;24:2343-64.

DeMaria CT, Sun Y, Long L, Wagner BJ, Brewer G. Structural determinants in AUF1 required for high affinity binding to A + U-rich elements. *J Biol Chem.* 1997;272:27635-43.

Denkert C, Weichert W, Winzer KJ et al. Expression of the ELAV-like protein HuR is associated with higher tumor grade and increased cyclooxygenase-2 expression in human breast carcinoma. *Clin Cancer Res.* 2004;10:5580-6.

Di Conza G, Mancini F, Buttarelli M, Pontecorvi A, Trimarchi F, Moretti F. MDM4 enhances p53 stability by promoting an active conformation of the protein upon DNA damage. *Cell Cycle.* 2012;11:749-60.

Didiano D, Hobert O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat Struct Mol Biol.* 2006;13:849-51.

Ding L, Ellis MJ, Li S et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature.* 2010;464:999-1005.

Dohi O, Yasui K, Gen Y et al. Epigenetic silencing of miR-335 and its host gene MEST in hepatocellular carcinoma. *Int J Oncol*. 2013;42:411-8.

Dredge BK, Darnell RB. Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. *Mol Cell Biol*. 2003;23:4687-700.

Dredge BK, Stefani G, Engelhard CC, Darnell RB. Nova autoregulation reveals dual functions in neuronal splicing. *EMBO J*. 2005;24:1608-20.

Du H, Cline MS, Osborne RJ et al. Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. *Nat Struct Mol Biol*. 2010;17:187-93.

Du L, Schageman JJ, Subauste MC et al. miR-93, miR-98, and miR-197 regulate expression of tumor suppressor gene FUS1. *Mol Cancer Res*. 2009;7:1234-43.

Dutertre M, Lacroix-Triki M, Driouch K et al. Exon-based clustering of murine breast tumor transcriptomes reveals alternative exons whose expression is associated with metastasis. *Cancer Res*. 2010;70:896-905.

Early P, Rogers J, Davis M et al. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell*. 1980;20:313-9.

Ebert MS, Sharp PA. Roles for microRNAs in conferring robustness to biological processes. *Cell*. 2012;149:515-24.

Eom T, Zhang C, Wang H et al. NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *eLife*. 2013;2:e00178-e00178.

Eulalio A, Rehwinkel J, Stricker M et al. Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. *Genes Dev*. 2007;21:2558-70.

Ezponda T, Pajares MJ, Agorreta J et al. The oncoprotein SF2/ASF promotes non-small cell lung cancer survival by enhancing survivin expression. *Clin Cancer Res*. 2010;16:4113-25.

Fan XC, Steitz JA. Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *EMBO J*. 1998;17:3448-60.

Fang L, Du WW, Yang X et al. Versican 3'-untranslated region (3'-UTR) functions as a ceRNA in inducing the development of hepatocellular carcinoma by regulating miRNA activity. *FASEB J*. 2013a;27:907-19.

Fang M, Simeonova I, Bardot B et al. Mdm4 loss in mice expressing a p53 hypomorph alters tumor spectrum without improving survival. *Oncogene*. 2013b

Fedele M, Fusco A. HMGA and cancer. *Biochim Biophys Acta*. 2010;1799:48-54.

Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*. 2008;9:102-14.

Fishelson Z, Donin N, Zell S, Schultz S, Kirschfink M. Obstacles to cancer immunotherapy: expression of membrane complement regulatory proteins (mCRPs) in tumors. *Mol Immunol*. 2003;40:109-23.

Forsman CL, Ng BC, Heinze RK et al. BMP-binding protein twisted gastrulation is required in mammary gland epithelium for normal ductal elongation and myoepithelial compartmentalization. *Dev Biol*. 2013;373:95-106.

Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A*. 2005;102:16176-81.

Fushimi K, Ray P, Kar A, Wang L, Sutherland LC, Wu JY. Up-regulation of the proapoptotic caspase 2 splicing isoform by a candidate tumor suppressor, RBM5. *Proc Natl Acad Sci U S A*. 2008;105:15708-13.

Gardina PJ, Clark TA, Shimada B et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*. 2006;7:325.

Gerson KD, Shearstone JR, Maddula VS, Seligmann BE, Mercurio AM. Integrin beta4 regulates SPARC protein to promote invasion. *J Biol Chem*. 2012;287:9835-44.

Geurts JM, Schoenmakers EF, Van de Ven WJ. Molecular characterization of a complex chromosomal rearrangement in a pleomorphic salivary gland adenoma involving the 3'-UTR of HMGIC. *Cancer Genet Cytogenet*. 1997;95:198-205.

Goncalves V, Matos P, Jordan P. Antagonistic SR proteins regulate alternative splicing of tumor-related Rac1b downstream of the PI3-kinase and Wnt pathways. *Hum Mol Genet*. 2009;18:3696-707.

Gonzalez-Aguilera C, Tous C, Gomez-Gonzalez B, Huertas P, Luna R, Aguilera A. The THP1-SAC3-SUS1-CDC31 complex works in transcription elongation-mRNA export preventing RNA-mediated genome instability. *Mol Biol Cell*. 2008;19:4310-8.

Gonzalez-Sancho JM, Aguilera O, Garcia JM et al. The Wnt antagonist DICKKOPF-1 gene is a downstream target of beta-catenin/TCF and is downregulated in human colon cancer. *Oncogene*. 2005;24:1098-103.

Goodarzi H, Najafabadi HS, Oikonomou P et al. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*. 2012;485:264-8.

Goujon M, McWilliam H, Li W et al. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*. 2010;38:W695-9.

Gout S, Brambilla E, Boudria A et al. Abnormal expression of the pre-mRNA splicing regulators SRSF1, SRSF2, SRPK1 and SRPK2 in non small cell lung carcinoma. *PLoS One*. 2012;7:e46539.

Grandinetti KB, Spengler BA, Biedler JL, Ross RA. Loss of one HuD allele on chromosome #1p selects for amplification of the N-myc proto-oncogene in human neuroblastoma cells. *Oncogene*. 2006;25:706-12.

Green RE, Lewis BP, Hillman RT et al. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics*. 2003;19 Suppl 1:i118-21.

Gregory RI, Yan KP, Amuthan G et al. The Microprocessor complex mediates the genesis of microRNAs. *Nature*. 2004;432:235-40.

Guil S, Caceres JF. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat Struct Mol Biol*. 2007;14:591-6.

Gunthert U, Hofmann M, Rudy W et al. A new variant of glycoprotein CD44 confers metastatic potential to rat carcinoma cells. *Cell*. 1991;65:13-24.

Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*. 2010;466:835-40.

Guo L, Chen C, Shi M et al. Stat3-coordinated Lin-28-let-7-HMGA2 and miR-200-ZEB1 circuits initiate and maintain oncostatin M-driven epithelial-mesenchymal transition. *Oncogene*. 2013

Guo L, Lu Z. The fate of miRNA* strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule? *PLoS One*. 2010;5:e11387.

Gyorffy B, Lanczky A, Eklund AC et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat*. 2010;123:725-31.

Ha I, Wightman B, Ruvkun G. A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient formation. *Genes Dev*. 1996;10:3041-50.

Hafner M, Landthaler M, Burger L et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*. 2010;141:129-41.

Hall CL, Bafico A, Dai J, Aaronson SA, Keller ET. Prostate cancer cells promote osteoblastic bone metastases through Wnts. *Cancer Res*. 2005;65:7554-60.

Hatzis C, Pusztai L, Valero V et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*. 2011;305:1873-81.

Hebert C, Norris K, Scheper MA, Nikitakis N, Sauk JJ. High mobility group A2 is a target for miRNA-98 in head and neck squamous cell carcinoma. *Mol Cancer*. 2007;6:5.

Heinonen M, Bono P, Narko K et al. Cytoplasmic HuR expression is a prognostic factor in invasive ductal breast carcinoma. *Cancer Res*. 2005;65:2157-61.

Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell*. 2013;153:654-65.

Hendrickson DG, Hogan DJ, McCullough HL et al. Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol*. 2009;7:e1000238.

Hess KR, Varadhachary GR, Taylor SH et al. Metastatic patterns in adenocarcinoma. *Cancer*. 2006;106:1624-33.

Heyn H, Engelmann M, Schreek S et al. MicroRNA miR-335 is crucial for the BRCA1 regulatory cascade in breast cancer development. *Int J Cancer*. 2011;129:2797-806.

Hill DA, Ivanovich J, Priest JR et al. DICER1 mutations in familial pleuropulmonary blastoma. *Science*. 2009;325:965.

Hillman-Jackson J, Clements D, Blankenberg D, Taylor J, Nekrutenko A. Using Galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics*. 2012;Chapter 10:Unit10.5.

Hoell JI, Larsson E, Runge S et al. RNA targets of wild-type and mutant FET family proteins. *Nat Struct Mol Biol*. 2011;18:1428-31.

Hoffmann I, Balling R. Cloning and expression analysis of a novel mesodermally expressed cadherin. *Dev Biol*. 1995;169:337-46.

Hung J, Kishimoto Y, Sugio K et al. Allele-specific chromosome 3p deletions occur at an early stage in the pathogenesis of lung carcinoma. *JAMA*. 1995;273:1908.

Huntzinger E, Kashima I, Fauser M, Sauliere J, Izaurralde E. SMG6 is the catalytic endonuclease that cleaves mRNAs containing nonsense codons in metazoan. *RNA*. 2008;14:2609-17.

Husemann Y, Geigl JB, Schubert F et al. Systemic spread is an early step in breast cancer. *Cancer Cell*. 2008;13:58-68.

Hwang HW, Wentzel EA, Mendell JT. Cell-cell contact globally activates microRNA biogenesis. *Proc Natl Acad Sci U S A*. 2009;106:7016-21.

Ibrahim F, Maragkakis M, Alexiou P, Maronski MA, Dichter MA, Mourelatos Z. Identification of in vivo, conserved, TAF15 RNA binding sites reveals the impact of TAF15 on the neuronal transcriptome. *Cell Rep*. 2013;3:301-8.

Ince-Dunn G, Okano HJ, Jensen KB et al. Neuronal Elav-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. *Neuron*. 2012;75:1067-80.

Iorio MV, Casalini P, Piovan C et al. microRNA-205 regulates HER3 in human breast cancer. *Cancer Res*. 2009;69:2195-200.

Ishigaki S, Masuda A, Fujioka Y et al. Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. *Sci Rep*. 2012;2:529.

Jacobsen A, Wen J, Marks DS, Krogh A. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Res*. 2010;20:1010-9.

Jani D, Lutz S, Marshall NJ et al. Sus1, Cdc31, and the Sac3 CID region form a conserved interaction platform that promotes nuclear pore association and mRNA export. *Mol Cell*. 2009;33:727-37.

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res*. 2011;39:e141.

Jensen KB, Darnell RB. CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. *Methods Mol Biol*. 2008;488:85-98.

Jensen KB, Dredge BK, Stefani G et al. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*. 2000;25:359-71.

Ji L, Nishizaki M, Gao B et al. Expression of several genes in the human chromosome 3p21.3 homozygous deletion region by an adenovirus vector results in tumor suppressor activities in vitro and in vivo. *Cancer Res*. 2002;62:2715-20.

Ji Z, Luo W, Li W et al. Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol Syst Biol*. 2011;7:534.

Jiang S, Zhang HW, Lu MH et al. MicroRNA-155 functions as an OncomiR in breast cancer by targeting the suppressor of cytokine signaling 1 gene. *Cancer Res*. 2010;70:3119-27.

Jin Y, Suzuki H, Maegawa S et al. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J*. 2003;22:905-12.

Kaddar T, Rouault JP, Chien WW et al. Two new miR-16 targets: caprin-1 and HMGA1, proteins implicated in cell proliferation. *Biol Cell*. 2009;101:511-24.

Kang Y, Siegel PM, Shu W et al. A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell*. 2003;3:537-49.

Kapinas K, Kessler C, Ricks T, Gronowicz G, Delany AM. miR-29 modulates Wnt signaling in human osteoblasts through a positive feedback loop. *J Biol Chem*. 2010;285:25221-31.

Kapinas K, Kessler CB, Delany AM. miR-29 suppression of osteonectin in osteoblasts: regulation during differentiation and by canonical Wnt signaling. *J Cell Biochem*. 2009;108:216-24.

Karam R, Carvalho J, Bruno I et al. The NMD mRNA surveillance pathway downregulates aberrant E-cadherin transcripts in gastric cancer cells and in CDH1 mutation carriers. *Oncogene*. 2008;27:4255-60.

Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol*. 2007;14:185-93.

Karni R, Hippo Y, Lowe SW, Krainer AR. The splicing-factor oncoprotein SF2/ASF activates mTORC1. *Proc Natl Acad Sci U S A*. 2008;105:15323-7.

Kashima I, Yamashita A, Izumi N et al. Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes Dev*. 2006;20:355-67.

Kedde M, Strasser MJ, Boldajipour B et al. RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell*. 2007;131:1273-86.

Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JA, Elkon R, Agami R. A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol*. 2010;12:1014-20.

Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc*. 2006;1:302-7.

Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*. 2010;11:345-55.

Khan AA, Betel D, Miller ML, Sander C, Leslie CS, Marks DS. Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nat Biotechnol*. 2009;27:549-55.

Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. *Cell*. 2003;115:209-16.

Kim HH, Kuwano Y, Srikantan S, Lee EK, Martindale JL, Gorospe M. HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev*. 2009a;23:1743-8.

Kim MY, Oskarsson T, Acharyya S et al. Tumor self-seeding by circulating cancer cells. *Cell*. 2009b;139:1315-26.

Kirschbaum-Slager N, Lopes GM, Galante PA, Riggins GJ, de Souza SJ. Splicing factors are differentially expressed in tumors. *Genet Mol Res*. 2004;3:512-20.

Klemke M, Meyer A, Hashemi Nezhad M, Belge G, Bartnitzke S, Bullerdiek J. Loss of let-7 binding sites resulting from truncations of the 3' untranslated region of HMGA2 mRNA in uterine leiomyomas. *Cancer Genet Cytogenet.* 2010;196:119-23.

Kole R, Krainer AR, Altman S. RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat Rev Drug Discov.* 2012;11:125-40.

Kondo M, Ji L, Kamibayashi C et al. Overexpression of candidate tumor suppressor gene FUS1 isolated from the 3p21.3 homozygous deletion region leads to G1 arrest and growth inhibition of lung cancer cells. *Oncogene.* 2001;20:6258-62.

Kong W, He L, Coppola M et al. MicroRNA-155 Regulates Cell Survival, Growth, and Chemosensitivity by Targeting FOXO3a in Breast Cancer. *Journal of Biological Chemistry.* 2010;285:17869-79.

Konig J, Zarnack K, Rot G et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol.* 2010;17:909-15.

Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 2011;39:D152-7.

Krol J, Busskamp V, Markiewicz I et al. Characterizing light-regulated retinal microRNAs reveals rapid turnover as a common property of neuronal microRNAs. *Cell.* 2010;141:618-31.

Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics.* 2012;28:1184-5.

Kumar MS, Lu J, Mercer KL, Golub TR, Jacks T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet.* 2007;39:673-7.

Kumar MS, Pester RE, Chen CY et al. Dicer1 functions as a haploinsufficient tumor suppressor. *Genes Dev.* 2009;23:2700-4.

Kundu P, Fabian MR, Sonenberg N, Bhattacharyya SN, Filipowicz W. HuR protein attenuates miRNA-mediated repression by promoting

miRISC dissociation from the target RNA. *Nucleic Acids Res.* 2012;40:5088-100.

Kwanhian W, Lenze D, Alles J et al. MicroRNA-142 is mutated in about 20% of diffuse large B-cell lymphoma. *Cancer Med.* 2012;1:141-55.

Lagier-Tourenne C, Polymenidou M, Hutt KR et al. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci.* 2012;15:1488-97.

Lagrange B, Martin RZ, Droin N et al. A role for miR-142-3p in colony-stimulating factor 1-induced monocyte differentiation into macrophages. *Biochim Biophys Acta.* 2013

Lambertz I, Nittner D, Mestdagh P et al. Monoallelic but not biallelic loss of Dicer1 promotes tumorigenesis in vivo. *Cell Death Differ.* 2010;17:633-41.

Landgraf P, Rusu M, Sheridan R et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell.* 2007;129:1401-14.

Lapuk A, Marr H, Jakkula L et al. Exon-level microarray analyses identify alternative splicing programs in breast cancer. *Mol Cancer Res.* 2010;8:961-74.

Lazarova DL, Spengler BA, Biedler JL, Ross RA. HuD, a neuronal-specific RNA-binding protein, is a putative regulator of N-myc pre-mRNA processing/stability in malignant human neuroblasts. *Oncogene.* 1999;18:2703-10.

Le Hir H, Gatfield D, Izaurralde E, Moore MJ. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.* 2001;20:4987-97.

Le Hir H, Moore MJ, Maquat LE. Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. *Genes Dev.* 2000;14:1098-108.

Lebedeva S, Jens M, Theil K et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell.* 2011;43:340-52.

Lee Y, Gamazon ER, Rebman E et al. Variants affecting exon skipping contribute to complex traits. *PLoS Genet.* 2012;8:e1002998.

Lee YS, Dutta A. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev.* 2007;21:1025-30.

Lejeune F, Li X, Maquat LE. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylation, and exonucleolytic activities. *Mol Cell.* 2003;12:675-87.

Lerman MI, Minna JD. The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium. *Cancer Res.* 2000;60:6116-33.

Leung AK, Young AG, Bhutkar A et al. Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat Struct Mol Biol.* 2011;18:237-44.

Levin JZ, Yassour M, Adiconis X et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods.* 2010;7:709-15.

Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A.* 2003;100:189-92.

Li F, Glinskii OV, Zhou J et al. Identification and analysis of signaling networks potentially involved in breast carcinoma metastasis to the brain. *PLoS One.* 2011a;6:e21977.

Li F, Glinskii OV, Zhou J et al. Identification and Analysis of Signaling Networks Potentially Involved in Breast Carcinoma Metastasis to the Brain. *PLoS ONE.* 2011b;6:e21977.

Li Y, Guo Z, Chen H et al. HOXC8-Dependent Cadherin 11 Expression Facilitates Breast Cancer Cell Migration through Trio and Rac. *Genes Cancer.* 2011c;2:880-8.

Li Y, Hong F, Yu Z. Decreased expression of microRNA-206 in breast cancer and its association with disease characteristics and patient survival. *J Int Med Res.* 2013

Lian WX, Yin RH, Kong XZ et al. THAP11, a novel binding protein of PCBP1, negatively regulates CD44 alternative splicing and cell invasion in a human hepatoma cell line. *FEBS Lett.* 2012;586:1431-8.

Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet.* 2010;11:75-87.

Licatalosi DD, Mele A, Fak JJ et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature.* 2008;456:464-9.

Licatalosi DD, Yano M, Fak JJ et al. Ptbp2 represses adult-specific splicing to regulate the generation of neuronal precursors in the embryonic brain. *Genes Dev.* 2012;26:1626-42.

Lieberman AP, Friedlich DL, Harmison G et al. Androgens regulate the mammalian homologues of invertebrate sex determination genes *tra-2* and *fox-1*. *Biochem Biophys Res Commun.* 2001;282:499-506.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A.* 2011;108:11093-8.

Lim LP, Lau NC, Garrett-Engele P et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature.* 2005;433:769-73.

Lin J, Xu K, Gitanjali J, Roth JA, Ji L. Regulation of tumor suppressor gene FUS1 expression by the untranslated regions of mRNA in human lung cancer cells. *Biochem Biophys Res Commun.* 2011;410:235-41.

Lin X, Miller JW, Mankodi A et al. Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. *Hum Mol Genet.* 2006;15:2087-97.

Ling S, Birnbaum Y, Nanhwan MK et al. Dickkopf-1 (DKK1) phosphatase and tensin homolog on chromosome 10 (PTEN) crosstalk via microRNA interference in the diabetic heart. *Basic Res Cardiol.* 2013;108:352.

Liu T, Tang H, Lang Y, Liu M, Li X. MicroRNA-27a functions as an oncogene in gastric adenocarcinoma by targeting prohibitin. *Cancer Lett.* 2009;273:233-42.

Liu WM, Guerra-Vladusic FK, Kurakata S, Lupu R, Kohwi-Shigematsu T. HMG-I(Y) recognizes base-unpairing regions of matrix attachment sequences and its increased expression is directly linked to metastatic breast cancer phenotype. *Cancer Res.* 1999;59:5695-703.

Liu X, Wang A, Heidbreder CE et al. MicroRNA-24 targeting RNA-binding protein DND1 in tongue squamous cell carcinoma. *FEBS Lett.* 2010;584:4115-20.

Loeb GB, Khan AA, Canner D et al. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol Cell.* 2012;48:760-70.

Lopez de Silanes I, Fan J, Yang X et al. Role of the RNA-binding protein HuR in colon carcinogenesis. *Oncogene.* 2003;22:7146-54.

Lu C, Stewart DJ, Lee JJ et al. Phase I clinical trial of systemically administered TUSC2(FUS1)-nanoparticles mediating functional gene transfer in humans. *PLoS One.* 2012;7:e34833.

Lu J, Getz G, Miska EA et al. MicroRNA expression profiles classify human cancers. *Nature.* 2005;435:834-8.

Lujambio A, Lowe SW. The microcosmos of cancer. *Nature.* 2012;482:347-55.

Luque FA, Furneaux HM, Ferziger R et al. Anti-Ri: an antibody associated with paraneoplastic opsoclonus and breast cancer. *Ann Neurol.* 1991;29:241-51.

Luzzi KJ, MacDonald IC, Schmidt EE et al. Multistep nature of metastatic inefficiency: dormancy of solitary cells after successful extravasation and limited survival of early micrometastases. *Am J Pathol.* 1998;153:865-73.

Lynch J, Fay J, Meehan M et al. MiRNA-335 suppresses neuroblastoma cell invasiveness by direct targeting of multiple genes from the non-canonical TGF-beta signalling pathway. *Carcinogenesis.* 2012;33:976-85.

Ma J, Flemr M, Stein P et al. MicroRNA activity is suppressed in mouse oocytes. *Curr Biol.* 2010a;20:265-70.

Ma L, Reinhardt F, Pan E et al. Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nat Biotechnol.* 2010b;28:341-7.

Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature.* 2007;449:682-8.

Malanchi I, Peinado H, Kassen D et al. Cutaneous cancer stem cell maintenance is dependent on beta-catenin signalling. *Nature.* 2008;452:650-3.

Malanchi I, Santamaria-Martinez A, Susanto E et al. Interactions between cancer stem cells and their niche govern metastatic colonization. *Nature.* 2012;481:85-9.

Mancini F, Di Conza G, Moretti F. MDM4 (MDMX) and its Transcript Variants. *Curr Genomics.* 2009;10:42-50.

Manjeshwar S, Branam DE, Lerner MR, Brackett DJ, Jupe ER. Tumor suppression by the prohibitin gene 3'untranslated region RNA in human breast cancer. *Cancer Res.* 2003;63:5251-6.

Manley GT, Smitt PS, Dalmau J, Posner JB. Hu antigens: reactivity with Hu antibodies, tumor expression, and major immunogenic sites. *Ann Neurol.* 1995;38:102-10.

Manley JL, Krainer AR. A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins).[letter]. *Genes Dev* 2010;24(11):1073-4.

Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* 2005;272:2118-31.

Martin EC, Elliott S, Rhodes LV et al. Preferential star strand biogenesis of pre-miR-24-2 targets PKC-alpha and suppresses cell survival in MCF-7 breast cancer cells. *Mol Carcinog.* 2012;n/a-n/a.

Masuda A, Andersen HS, Doktor TK et al. CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. *Sci Rep.* 2012;2:209.

Matter N, Herrlich P, König H. Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature*. 2002;420:691-5.

Mattiske S, Suetani RJ, Neilsen PM, Callen DF. The oncogenic role of miR-155 in breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2012;21:1236-43.

Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*. 2009;138:673-84.

Mayr C, Hemann MT, Bartel DP. Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science*. 2007;315:1576-9.

Mazan-Mamczarz K, Hagner PR, Corl S et al. Post-transcriptional gene regulation by HuR promotes a more tumorigenic phenotype. *Oncogene*. 2008a;27:6151-63.

Mazan-Mamczarz K, Hagner PR, Dai B et al. Identification of transformation-related pathways in a breast epithelial cell model using a ribonomics approach. *Cancer Res*. 2008b;68:7730-5.

Mehta A, Kinter MT, Sherman NE, Driscoll DM. Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol Cell Biol*. 2000;20:1846-54.

Meindl A, Ditsch N, Kast K, Rhiem K, Schmutzler RK. Hereditary breast and ovarian cancer: new genes, new treatments, new concepts. *Dtsch Arztebl Int*. 2011;108:323-30.

Mendoza M, Khanna C. Revisiting the seed and soil in cancer metastasis. *Int J Biochem Cell Biol*. 2009;41:1452-62.

Menezes ME, Devine DJ, Shevde LA, Samant RS. Dickkopf1: a tumor suppressor or metastasis promoter? *Int J Cancer*. 2012;130:1477-83.

Meri S, Morgan BP, Davies A et al. Human protectin (CD59), an 18,000-20,000 MW complement lysis restricting factor, inhibits C5b-8 catalysed insertion of C9 into lipid bilayers. *Immunology*. 1990;71:1-9.

Mertens-Talcott SU, Chintharlapalli S, Li X, Safe S. The oncogenic microRNA-27a targets genes that regulate specificity protein transcription

factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells. *Cancer Res.* 2007;67:11001-11.

Michlewski G, Guil S, Caceres JF. Stimulation of pri-miR-18a Processing by hnRNP A1. *Adv Exp Med Biol.* 2010;700:28-35.

Middendorp S, Paoletti A, Schiebel E, Bornens M. Identification of a new mammalian centrin gene, more closely related to *Saccharomyces cerevisiae* CDC31 gene. *Proc Natl Acad Sci U S A.* 1997;94:9141-6.

Mili S, Steitz JA. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA.* 2004;10:1692-4.

Miller JW, Urbinati CR, Teng-Umnuy P et al. Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J.* 2000;19:4439-48.

Minn AJ, Gupta GP, Padua D et al. Lung metastasis genes couple breast tumor size and metastatic spread. *Proc Natl Acad Sci U S A.* 2007;104:6740-5.

Minn AJ, Gupta GP, Siegel PM et al. Genes that mediate breast cancer metastasis to lung. *Nature.* 2005a;436:518-24.

Minn AJ, Kang Y, Serganova I et al. Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J Clin Invest.* 2005b;115:44-55.

Mishima Y, Giraldez AJ, Takeda Y et al. Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. *Curr Biol.* 2006;16:2135-42.

Missiaglia E, Shepherd CJ, Patel S et al. MicroRNA-206 expression levels correlate with clinical behaviour of rhabdomyosarcomas. *Br J Cancer.* 2010;102:1769-77.

Mitra A, Menezes ME, Shevde LA, Samant RS. DNAJB6 induces degradation of beta-catenin and causes partial reversal of mesenchymal phenotype. *J Biol Chem.* 2010;285:24686-94.

Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*. 2009;136:688-700.

Morgenstern JP, Land H. Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Res*. 1990;18:3587-96.

Morlando M, Dini Modigliani S, Torrelli G et al. FUS stimulates microRNA biogenesis by facilitating co-transcriptional Drosha recruitment. *EMBO J*. 2012;31:4502-10.

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat*. 2008;29:1037-47.

Mourtada-Maarabouni M, Sutherland LC, Meredith JM, Williams GT. Simultaneous acceleration of the cell cycle and suppression of apoptosis by splice variant delta-6 of the candidate tumour suppressor LUCA-15/RBM5. *Genes Cells*. 2003;8:109-19.

Mukherjee N, Corcoran DL, Nusbaum JD et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol Cell*. 2011;43:327-39.

Mukherji S, Ebert MS, Zheng GX, Tsang JS, Sharp PA, van Oudenaarden A. MicroRNAs can generate thresholds in target gene expression. *Nat Genet*. 2011;43:854-9.

Nakaya T, Alexiou P, Maragkakis M, Chang A, Mourelatos Z. FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns. *RNA*. 2013;19:498-509.

Narita M, Narita M, Krizhanovsky V et al. A novel role for high-mobility group a proteins in cellular senescence and heterochromatin formation. *Cell*. 2006;126:503-14.

Nelson PT, De Planell-Saguer M, Lamprinaki S et al. A novel monoclonal antibody against human Argonaute proteins reveals unexpected characteristics of miRNAs in human blood cells. *RNA*. 2007;13:1787-92.

Nguyen DX, Bos PD, Massague J. Metastasis: from dissemination to organ-specific colonization. *Nat Rev Cancer*. 2009;9:274-84.

Niehrs C. Function and biological roles of the Dickkopf family of Wnt modulators. *Oncogene*. 2006;25:7469-81.

Nolde MJ, Saka N, Reinert KL, Slack FJ. The *Caenorhabditis elegans* pumilio homolog, *puf-9*, is required for the 3'UTR-mediated repression of the *let-7* microRNA target gene, *hbl-1*. *Dev Biol*. 2007;305:551-63.

Oh JJ, Razfar A, Delgado I et al. 3p21.3 tumor suppressor gene H37/Luca15/RBM5 inhibits growth of human lung cancer cells through cell cycle arrest and apoptosis. *Cancer Res*. 2006;66:3419-27.

Oh JJ, West AR, Fishbein MC, Slamon DJ. A candidate tumor suppressor gene, H37, from the human lung cancer tumor suppressor locus 3p21.3. *Cancer Res*. 2002;62:3207-13.

Ohgaki K, Iida A, Kasumi F et al. Mapping of a new target region of allelic loss to a 6-cM interval at 21q21 in primary breast cancers. *Genes Chromosomes Cancer*. 1998;23:244-7.

Okano HJ, Darnell RB. A hierarchy of Hu RNA binding proteins in developing and adult neurons. *J Neurosci*. 1997;17:3024-37.

Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*. 2009;4:14.

Oskarsson T, Acharyya S, Zhang XH et al. Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs. *Nat Med*. 2011;17:867-74.

Palmieri D, D'Angelo D, Valentino T et al. Downregulation of HMGA-targeting microRNAs has a critical role in human pituitary tumorigenesis. *Oncogene*. 2012;31:3857-65.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. 2008;40:1413-5.

Pandit S, Zhou Y, Shiue L et al. Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. *Mol Cell*. 2013;50:223-35.

Paoletti A, Bordes N, Haddad R, Schwartz CL, Chang F, Bornens M. Fission yeast cdc31p is a component of the half-bridge and controls SPB duplication. *Mol Biol Cell*. 2003;14:2793-808.

Paronetto MP, Minana B, Valcarcel J. The Ewing sarcoma protein regulates DNA damage-induced alternative splicing. *Mol Cell*. 2011;43:353-68.

Patanaphan V, Salazar OM, Risco R. Breast cancer: metastatic patterns and their prognosis. *South Med J*. 1988;81:1109-12.

Pencheva N, Tran H, Buss C et al. Convergent multi-miRNA targeting of ApoE drives LRP1/LRP8-dependent melanoma metastasis and angiogenesis. *Cell*. 2012;151:1068-82.

Perrin-Vidoz L, Sinilnikova OM, Stoppa-Lyonnet D, Lenoir GM, Mazoyer S. The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons. *Hum Mol Genet*. 2002;11:2805-14.

Petitjean A, Mathe E, Kato S et al. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum Mutat*. 2007;28:622-9.

Pinzone JJ, Hall BM, Thudi NK et al. The role of Dickkopf-1 in bone development, homeostasis, and disease. *Blood*. 2009;113:517-25.

Pishvaian MJ, Feltes CM, Thompson P, Bussemakers MJ, Schalken JA, Byers SW. Cadherin-11 is expressed in invasive breast cancer cell lines. *Cancer Res*. 1999;59:947-52.

Png KJ, Halberg N, Yoshida M, Tavazoie SF. A microRNA regulon that mediates endothelial recruitment and metastasis by cancer cells. *Nature*. 2012;481:190-4.

Png KJ, Yoshida M, Zhang XH et al. MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer. *Genes Dev.* 2011;25:226-31.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature.* 2010;465:1033-8.

Polymenidou M, Lagier-Tourenne C, Hutt KR et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat Neurosci.* 2011;14:459-68.

Polytarchou C, Iliopoulos D, Struhl K. An integrated transcriptional regulatory circuit that reinforces the breast cancer stem cell state. *Proc Natl Acad Sci U S A.* 2012;109:14470-5.

Ponta H, Sherman L, Herrlich PA. CD44: from adhesion molecules to signalling regulators. *Nat Rev Mol Cell Biol.* 2003;4:33-45.

Qiao L, Xu ZL, Zhao TJ, Ye LH, Zhang XD. Dkk-1 secreted by mesenchymal stem cells inhibits growth of breast cancer cells via depression of Wnt signalling. *Cancer Lett.* 2008;269:67-77.

Qin Q, Xu Y, He T, Qin C, Xu J. Normal and disease-related biological functions of Twist1 and underlying molecular mechanisms. *Cell Res.* 2012;22:90-106.

Rahman L, Bliskovski V, Reinhold W, Zajac-Kaye M. Alternative splicing of brain-specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles. *Genomics.* 2002;80:245-9.

Rallapalli R, Strachan G, Cho B, Mercer WE, Hall DJ. A novel MDMX transcript expressed in a variety of transformed cell lines encodes a truncated protein with potent p53 repressive activity. *J Biol Chem.* 1999;274:8299-308.

Rallapalli R, Strachan G, Tuan RS, Hall DJ. Identification of a domain within MDMX-S that is responsible for its high affinity interaction with p53 and high-level expression in mammalian cells. *J Cell Biochem.* 2003;89:563-75.

Rastinejad F, Conboy MJ, Rando TA, Blau HM. Tumor suppression by RNA from the 3' untranslated region of alpha-tropomyosin. *Cell*. 1993;75:1107-17.

Reeves R, Edberg DD, Li Y. Architectural transcription factor HMGI(Y) promotes tumor progression and mesenchymal transition of human epithelial cells. *Mol Cell Biol*. 2001;21:575-94.

Reid DC, Chang BL, Gunderson SI, Alpert L, Thompson WA, Fairbrother WG. Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA*. 2009;15:2385-97.

Reinke LM, Xu Y, Cheng C. Snail Represses the Splicing Regulator Epithelial Splicing Regulatory Protein 1 to Promote Epithelial-Mesenchymal Transition. *Journal of Biological Chemistry*. 2012;287:36435-42.

Religio A, Ben-Dov C, Baum M et al. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem*. 2005;280:4779-84.

Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*. 2000;16:276-7.

Riley KJ, Rabinowitz GS, Yario TA, Luna JM, Darnell RB, Steitz JA. EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *EMBO J*. 2012;31:2207-21.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139-40.

Rogelj B, Easton LE, Bogu GK et al. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci Rep*. 2012;2:603.

Rosenfeld N, Aharonov R, Meiri E et al. MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol*. 2008;26:462-9.

Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer*. 2012;12:68-78.

Rutnam ZJ, Yang BB. The non-coding 3' UTR of CD44 induces metastasis by regulating extracellular matrix functions. *J Cell Sci.* 2012;125:2075-85.

Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell.* 2011;146:353-8.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science.* 2008;320:1643-7.

Sankar S, Lessnick SL. Promiscuous partnerships in Ewing's sarcoma. *Cancer Genet.* 2011;204:351-65.

Sanz DJ, Acedo A, Infante M et al. A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. *Clin Cancer Res.* 2010;16:1957-67.

Sarrio D, Rodriguez-Pinilla SM, Hardisson D, Cano A, Moreno-Bueno G, Palacios J. Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* 2008;68:989-97.

Scarola M, Schoeftner S, Schneider C, Benetti R. miR-335 directly targets Rb1 (pRb/p105) in a proximal connection to p53-dependent stress response. *Cancer Res.* 2010;70:6925-33.

Schild D, Ananthaswamy HN, Mortimer RK. An endomitotic effect of a cell cycle mutation of *Saccharomyces cerevisiae*. *Genetics.* 1981;97:551-62.

Schoeftner S, Scarola M, Comisso E, Schneider C, Benetti R. An Oct4-pRb axis, controlled by MiR-335, integrates stem cell self-renewal and cell cycle control. *Stem Cells.* 2012;31:717-28.

Schoenmakers EF, Wanschura S, Mols R, Bullerdiek J, Van den Berghe H, Van de Ven WJ. Recurrent rearrangements in the high mobility group protein gene, HMGI-C, in benign mesenchymal tumours. *Nat Genet.* 1995;10:436-44.

Schuijers J, Clevers H. Adult mammalian stem cells: the role of Wnt, Lgr5 and R-spondins. *EMBO J.* 2012;31:2685-96.

Scott GK, Mattie MD, Berger CE, Benz SC, Benz CC. Rapid alteration of microRNA levels by histone deacetylase inhibition. *Cancer Res.* 2006;66:1277-81.

Scott IC, Blitz IL, Pappano WN, Maas SA, Cho KW, Greenspan DS. Homologues of Twisted gastrulation are extracellular cofactors in antagonism of BMP signalling. *Nature.* 2001;410:475-8.

Seitz H. Redefining microRNA targets. *Curr Biol.* 2009;19:870-3.

Selbach M, Schwanh  usser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature.* 2008;455:58-63.

Sethi N, Kang Y. Unravelling the complexity of metastasis - molecular understanding and targeted therapies. *Nat Rev Cancer.* 2011;11:735-48.

Shah SN, Cope L, Poh W et al. HMGA1: A Master Regulator of Tumor Progression in Triple-Negative Breast Cancer Cells. *PLoS One.* 2013;8:e63419.

Shao C, Zhao L, Wang K, Xu W, Zhang J, Yang B. The tumor suppressor gene RBM5 inhibits lung adenocarcinoma cell growth and induces apoptosis. *World J Surg Oncol.* 2012;10:160.

Shapiro IM, Cheng AW, Flytzanis NC et al. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet.* 2011;7:e1002218.

Sharp PA. The centrality of RNA. *Cell.* 2009;136:577-80.

Shaw G, Kamen R. A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell.* 1986;46:659-67.

Shen WW, Zeng Z, Zhu WX, Fu GH. MiR-142-3p functions as a tumor suppressor by targeting CD133, ABCG2, and Lgr5 in colon cancer cells. *J Mol Med (Berl).* 2013

Shi L, Jiang D, Sun G et al. miR-335 promotes cell proliferation by directly targeting Rb1 in meningiomas. *J Neurooncol.* 2012;110:155-62.

Shu M, Zheng X, Wu S et al. Targeting oncogenic miR-335 inhibits growth and invasion of malignant astrocytoma cells. *Mol Cancer*. 2011;10:59.

Shvarts A, Steegenga WT, Riteco N et al. MDMX: a novel p53-binding protein with some functional properties of MDM2. *EMBO J*. 1996;15:5349-57.

Siegel R, DeSantis C, Virgo K et al. Cancer treatment and survivorship statistics, 2012. *CA Cancer J Clin*. 2012;62:220-41.

Simone LE, Keene JD. Mechanisms coordinating ELAV/Hu mRNA regulons. *Curr Opin Genet Dev*. 2013;23:35-43.

Singh A, Karnoub AE, Palmby TR, Lengyel E, Sondek J, Der CJ. Rac1b, a tumor associated, constitutively active Rac1 splice variant, promotes cellular transformation. *Oncogene*. 2004;23:9369-80.

Singh AM, Reynolds D, Cliff T et al. Signaling network crosstalk in human pluripotent cells: a Smad2/3-regulated switch that controls the balance between self-renewal and differentiation. *Cell Stem Cell*. 2012;10:312-26.

Sorrentino A, Liu CG, Addario A, Peschle C, Scambia G, Ferlini C. Role of microRNAs in drug-resistant ovarian cancer cells. *Gynecol Oncol*. 2008;111:478-86.

Spizzo R, Nicoloso MS, Croce CM, Calin GA. SnapShot: MicroRNAs in Cancer. *Cell*. 2009;137:586-586.e1.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res*. 2011;21:1563-71.

Suh N, Baehner L, Moltzahn F et al. MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr Biol*. 2010;20:271-7.

Sundaresan V, Ganly P, Hasleton P et al. p53 and chromosome 3 abnormalities, characteristic of malignant lung tumours, are detectable in preinvasive lesions of the bronchus. *Oncogene*. 1992;7:1989-97.

Suzuki H, Toyota M, Carraway H et al. Frequent epigenetic inactivation of Wnt antagonist genes in breast cancer. *Br J Cancer*. 2008;98:1147-56.

Suzuki HI, Yamagata K, Sugimoto K, Iwamoto T, Kato S, Miyazono K. Modulation of microRNA processing by p53. *Nature*. 2009;460:529-33.

Szabo A, Dalmau J, Manley G et al. HuD, a paraneoplastic encephalomyelitis antigen, contains RNA-binding domains and is homologous to Elav and Sex-lethal. *Cell*. 1991;67:325-33.

Takeo K, Kawai T, Nishida K et al. Oxidative stress-induced alternative splicing of transformer 2beta (SFRS10) and CD44 pre-mRNAs in gastric epithelial cells. *Am J Physiol Cell Physiol*. 2009;297:C330-8.

Tang J, Ahmad A, Sarkar FH. The Role of MicroRNAs in Breast Cancer Migration, Invasion and Metastasis. *Int J Mol Sci*. 2012a;13:13414-37.

Tang W, Zhu J, Su S et al. MiR-27 as a prognostic marker for breast cancer progression and patient survival. *PLoS One*. 2012b;7:e51702.

Tariq MA, Kim HJ, Jejelowo O, Pourmand N. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res*. 2011;39:e120.

Tavazoie SF, Alarcon C, Oskarsson T et al. Endogenous human microRNAs that suppress breast cancer metastasis. *Nature*. 2008;451:147-52.

Tay Y, Kats L, Salmena L et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*. 2011;147:344-57.

Thompson D, Easton D. The genetic epidemiology of breast cancer genes. *J Mammary Gland Biol Neoplasia*. 2004;9:221-36.

Thomson JM, Newman M, Parker JS, Morin-Kensicki EM, Wright T, Hammond SM. Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev*. 2006;20:2202-7.

Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*. 2005;33:201-12.

Timmer T, Terpstra P, van den Berg A et al. A comparison of genomic structures and expression patterns of two closely related flanking genes in a critical lung cancer region at 3p21.3. *Eur J Hum Genet.* 1999;7:478-86.

Tollervey JR, Curk T, Rogelj B et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci.* 2011;14:452-8.

Tome M, Lopez-Romero P, Albo C et al. miR-335 orchestrates cell proliferation, migration and differentiation in human mesenchymal stem cells. *Cell Death Differ.* 2011;18:985-95.

Tran H, Gourrier N, Lemerrier-Neuillet C et al. Analysis of exonic regions involved in nuclear localization, splicing activity, and dimerization of Muscleblind-like-1 isoforms. *J Biol Chem.* 2011;286:16435-46.

Trojan P, Krauss N, Choe HW, Giessl A, Pulvermuller A, Wolfrum U. Centrins in retinal photoreceptor cells: regulators in the connecting cilium. *Prog Retin Eye Res.* 2008;27:237-59.

Uhlen M, Oksvold P, Fagerberg L et al. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol.* 2010;28:1248-50.

Ule J, Jensen K, Mele A, Darnell RB. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods.* 2005a;37:376-86.

Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. *Science.* 2003;302:1212-5.

Ule J, Stefani G, Mele A et al. An RNA map predicting Nova-dependent splicing regulation. *Nature.* 2006;444:580-6.

Ule J, Ule A, Spencer J et al. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet.* 2005b;37:844-52.

Underwood JG, Boutz PL, Dougherty JD, Stoilov P, Black DL. Homologues of the *Caenorhabditis elegans* Fox-1 protein are neuronal splicing regulators in mammals. *Mol Cell Biol.* 2005;25:10005-16.

Vakalopoulou E, Schaack J, Shenk T. A 32-kilodalton protein binds to AU-rich domains in the 3' untranslated regions of rapidly degraded mRNAs. *Mol Cell Biol.* 1991;11:3355-64.

Valastyan S, Chang A, Benaich N, Reinhardt F, Weinberg RA. Activation of miR-31 function in already-established metastases elicits metastatic regression. *Genes Dev.* 2011;25:646-59.

Valastyan S, Reinhardt F, Benaich N et al. A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis. *Cell.* 2009;137:1032-46.

van 't Veer LJ, Dai H, van de Vijver MJ et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415:530-6.

van de Vijver MJ, He YD, van't Veer LJ et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002;347:1999-2009.

Vandesompele J, De Preter K, Pattyn F et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 2002;3:RESEARCH0034.

Varsano S, Frolkis I, Ophir D. Expression and distribution of cell-membrane complement regulatory glycoproteins along the human respiratory tract. *Am J Respir Crit Care Med.* 1995;152:1087-93.

Varsano S, Rashkovsky L, Shapiro H, Ophir D, Mark-Bentankur T. Human lung cancer cell lines express cell membrane complement inhibitory proteins and are extremely resistant to complement-mediated lysis; a comparison with normal human respiratory epithelium in vitro, and an insight into mechanism(s) of resistance. *Clin Exp Immunol.* 1998;113:173-82.

Vasudevan S. Posttranscriptional upregulation by microRNAs. *Wiley Interdiscip Rev RNA.* 2012;3:311-30.

Vella MC, Reinert K, Slack FJ. Architecture of a validated microRNA::target interaction. *Chem Biol.* 2004;11:1619-23.

Venables JP, Brosseau JP, Gadea G et al. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Mol Cell Biol.* 2013;33:396-405.

Venables JP, Klinck R, Bramard A et al. Identification of alternative splicing markers for breast cancer. *Cancer Res.* 2008;68:9525-31.

Venables JP, Klinck R, Koh C et al. Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol.* 2009;16:670-6.

Voorhoeve PM, le Sage C, Schrier M et al. A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Adv Exp Med Biol.* 2007;604:17-46.

Wagner BJ, DeMaria CT, Sun Y, Wilson GM, Brewer G. Structure and genomic organization of the human AUF1 gene: alternative pre-mRNA splicing generates four protein isoforms. *Genomics.* 1998;48:195-202.

Wahl MC, Will CL, Luhrmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell.* 2009;136:701-18.

Wang C, Gao C, Zhuang JL, Ding C, Wang Y. A combined approach identifies three mRNAs that are down-regulated by microRNA-29b and promote invasion ability in the breast cancer cell line MCF-7. *J Cancer Res Clin Oncol.* 2012;138:2127-36.

Wang D, Wengrod J, Gardner LB. Overexpression of the c-myc oncogene inhibits nonsense-mediated RNA decay in B lymphocytes. *J Biol Chem.* 2011a;286:40038-43.

Wang D, Zavadil J, Martin L et al. Inhibition of Nonsense-Mediated RNA Decay by the Tumor Microenvironment Promotes Tumorigenesis. *Molecular and Cellular Biology.* 2011b;31:3670-80.

Wang ET, Cody NA, Jog S et al. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell.* 2012;150:710-24.

Wang ET, Sandberg R, Luo S et al. Alternative isoform regulation in human tissue transcriptomes. *Nature.* 2008;456:470-6.

Wang L, Wang J. MicroRNA-mediated breast cancer metastasis: from primary site to distant organs. *Oncogene.* 2012;31:2499-511.

Wang X, Ling C, Bai Y, Zhao J. MicroRNA-206 is associated with invasion and metastasis of lung cancer. *Anat Rec (Hoboken).* 2011c;294:88-92.

Wappenschmidt B, Becker AA, Hauke J et al. Analysis of 30 putative BRCA1 splicing mutations in hereditary breast and ovarian cancer families identifies exonic splice site mutations that escape in silico prediction. *PLoS One.* 2012;7:e50800.

Warzecha CC, Jiang P, Amirikian K et al. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J.* 2010;29:3286-300.

Warzecha CC, Sato TK, Nabet B, Hogenesch JB, Carstens RP. ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol Cell.* 2009;33:591-601.

Watermann DO, Tang Y, Zur Hausen A, Jager M, Stamm S, Stickeler E. Splicing factor Tra2-beta1 is specifically induced in breast cancer and regulates alternative splicing of the CD44 gene. *Cancer Res.* 2006;66:4774-80.

Wei JJ, Wu X, Peng Y et al. Regulation of HMGA1 expression by microRNA-296 affects prostate cancer growth and invasion. *Clin Cancer Res.* 2011;17:1297-305.

Weidensdorfer D, Stohr N, Baude A et al. Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs. *RNA.* 2009;15:104-15.

Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell.* 1993;75:855-62.

Wistuba II, Behrens C, Milchgrub S et al. Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. *Oncogene.* 1999;18:643-50.

Wu C, Orozco C, Boyer J et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* 2009a;10:R130.

Wu H, Neilson JR, Kumar P et al. miRNA profiling of naive, effector and memory CD8 T cells. *PLoS One.* 2007;2:e1020.

Wu H, Zhu S, Mo YY. Suppression of cell growth and invasion by miR-205 in breast cancer. *Cell Res.* 2009b;19:439-48.

Wu L, Cai C, Wang X, Liu M, Li X, Tang H. MicroRNA-142-3p, a new regulator of RAC1, suppresses the migration and invasion of hepatocellular carcinoma cells. *FEBS Lett.* 2011a;585:1322-30.

Wu X, Brewer G. The regulation of mRNA stability in mammalian cells: 2.0. *Gene*. 2012;500:10-21.

Wu X, Shen QT, Oristian DS et al. Skin stem cells orchestrate directional migration by regulating microtubule-ACF7 connections through GSK3beta. *Cell*. 2011b;144:341-52.

Xu Y, Zhao F, Wang Z et al. MicroRNA-335 acts as a metastasis suppressor in gastric cancer by targeting Bcl-w and specificity protein 1. *Oncogene*. 2012;31:1398-407.

Xue Y, Zhou Y, Wu T et al. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*. 2009;36:996-1006.

Yae T, Tsuchihashi K, Ishimoto T et al. Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. *Nat Commun*. 2012;3:883.

Yamada H, Yanagisawa K, Tokumaru S et al. Detailed characterization of a homozygously deleted region corresponding to a candidate tumor suppressor locus at 21q11-21 in human lung cancer. *Genes Chromosomes Cancer*. 2008;47:810-8.

Yamagata K, Fujiyama S, Ito S et al. Maturation of microRNA is hormonally regulated by a nuclear receptor. *Mol Cell*. 2009;36:340-7.

Yan G, Fukabori Y, McBride G, Nikolaropolous S, McKeehan WL. Exon switching and activation of stromal and embryonic fibroblast growth factor (FGF)-FGF receptor genes in prostate epithelial cells accompany stromal independence and malignancy. *Mol Cell Biol*. 1993;13:4513-22.

Yan J, Marr TG. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res*. 2005;15:369-75.

Yang JS, Phillips MD, Betel D et al. Widespread regulatory activity of vertebrate microRNA* species. *RNA*. 2011;17:312-26.

Yeo GW, Coufal NG, Liang TY, Peng GE, Fu XD, Gage FH. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*. 2009;16:130-7.

Yilmaz M, Christofori G. EMT, the cytoskeleton, and cancer cell invasion. *Cancer Metastasis Rev.* 2009;28:15-33.

Yoneda T, Williams PJ, Hiraga T, Niewolna M, Nishimura R. A Bone-Seeking Clone Exhibits Different Biological Properties from the MDA-MB-231 Parental Human Breast Cancer Cells and a Brain-Seeking Clone In Vivo and In Vitro. *J Bone Miner Res.* 2001;16:1486-95.

Yoshida K, Sanada M, Shiraishi Y et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature.* 2011;478:64-9.

Young LE, Moore AE, Sokol L, Meisner-Kober N, Dixon DA. The mRNA stability factor HuR inhibits microRNA-16 targeting of COX-2. *Mol Cancer Res.* 2012;10:167-80.

Youngren KK, Coveney D, Peng X et al. The Ter mutation in the dead end gene causes germ cell loss and testicular germ cell tumours. *Nature.* 2005;435:360-4.

Zabarovsky ER, Lerman MI, Minna JD. Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene.* 2002;21:6915-35.

Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol.* 2011;29:607-14.

Zhang C, Frias MA, Mele A et al. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science.* 2010a;329:439-43.

Zhang C, Zhang Z, Castle J et al. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.* 2008;22:2550-63.

Zhang J, Tu Q, Bonewald LF et al. Effects of miR-335-5p in modulating osteogenic differentiation by specifically downregulating Wnt antagonist DKK1. *J Bone Miner Res.* 2011;26:1953-63.

Zhang L, Huang J, Yang N et al. microRNAs exhibit high frequency genomic alterations in human cancer. *Proc Natl Acad Sci U S A.* 2006;103:9136-41.

Zhang T, Huang X-H, Dong L et al. PCBP-1 regulates alternative splicing of the CD44 gene and inhibits invasion in human hepatoma cell line HepG2 cells. *Mol Cancer*. 2010b;9:72.

Zhang Y, Yang P, Sun T et al. miR-126 and miR-126* repress recruitment of mesenchymal stem cells and inflammatory monocytes to inhibit breast cancer metastasis. *Nat Cell Biol*. 2013;15:284-94.

Zhang Z, Xin D, Wang P et al. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol*. 2009;7:23.

Zhou AD, Diao LT, Xu H et al. beta-Catenin/LEF1 transactivates the microRNA-371-373 cluster that modulates the Wnt/beta-catenin-signaling pathway. *Oncogene*. 2012;31:2968-78.

Zhou X, Benson KF, Ashar HR, Chada K. Mutation responsible for the mouse pygmy phenotype in the developmentally regulated factor HMGI-C. *Nature*. 1995;376:771-4.

Zisoulis DG, Lovci MT, Wilbert ML et al. Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat Struct Mol Biol*. 2010;17:173-9.

Zou L, Zhang H, Du C et al. Correlation of SRSF1 and PRMT1 expression with clinical status of pediatric acute lymphoblastic leukemia. *J Hematol Oncol*. 2012;5:42.

Zou T, Rao JN, Liu L et al. Polyamines regulate the stability of JunD mRNA by modulating the competitive binding of its 3' untranslated region to HuR and AUF1. *Mol Cell Biol*. 2010;30:5021-32.

Zovoilis A, Smorag L, Pantazi A, Engel W. Members of the miR-290 cluster modulate in vitro differentiation of mouse embryonic stem cells. *Differentiation*. 2009;78:69-78.

Zucconi BE, Wilson GM. Modulation of neoplastic gene regulatory pathways by the RNA-binding factor AUF1. *Front Biosci*. 2011;16:2307-25.