

**A NOVEL EXPERIMENTAL APPROACH TO MAP MICRORNA-RNA
INTERACTIONS *IN VIVO***

by
Xiaoyi Li

A dissertation

Presented to the Faculty of the Louis V. Gerstner Jr.

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

New York, NY

May 2021

Andrea Ventura, MD, PhD

Dissertation Mentor

Date:

Copyright by Xiaoyi Li 2021

DEDICATION

I would like to dedicate this thesis to my parents, Zhong Li and Wen Han.

ABSTRACT

The identification of miRNA targets by Ago2 crosslinking-immunoprecipitation (CLIP) methods has provided major insights into the biology of this important class of non-coding RNAs. However, these methods are technically challenging and difficult to apply to an *in vivo* setting.

To overcome these limitations and to facilitate the investigation of miRNA functions in mice, we have developed a method (HEAP: for Halo-Enhanced Ago2 Pulldown) to map miRNA-RNA interaction sites. This method is based on a novel genetically engineered mouse harboring a conditional, Cre-regulated, Halo-Ago2 allele expressed from the endogenous *Ago2* locus. By using a resin conjugated to the HaloTag ligand, Ago2-miRNA-RNA complexes can be efficiently purified from cells and tissues expressing the endogenous Halo-Ago2 allele. We demonstrate the reproducibility and sensitivity of this method in mouse embryonic stem cells, in developing embryos, in adult tissues, and in autochthonous mouse models of human brain and lung cancers.

The tools and the datasets we have generated will serve as a valuable resource to the scientific community and will facilitate the characterization of miRNA functions under physiological and pathological conditions.

BIOGRAPHICAL SKETCH

Xiaoyi was born to Zhong Li and Wen Han on September 12th, 1991 in Guiyang, China. She graduated from Guiyang No.1 High School in 2010 and moved to Beijing, where she did her undergraduate study in the next 4 years. Xiaoyi majored in Biological Sciences in Tsinghua University and completed her thesis research on the central nervous system infiltration of T-cell acute lymphoblastic leukemia under the supervision of Dr. Wei Guo in the School of Medicine, Tsinghua university. In the summer of 2013, she flew to UK and spent 3 months in the laboratory of Dr. Ruth Murrell-Lagnado working on characterizing the store-operated calcium entry pathway. During her time in Tsinghua University, she was also a member of the Tsinghua Talented Program in Biological Sciences and was awarded Honors Student in 2013. She graduated with a Bachelor of Science (BS) in 2014 and was awarded the outstanding undergraduate student in Beijing. Immediately after her graduation, Xiaoyi moved to New York City and started as a PhD student at the Louis V. Gerstner Jr. Graduate School of Biomedical Sciences at Memorial Sloan Kettering Cancer Center. She joined the laboratory of Dr. Andrea Ventura in the summer of 2015 and has focused her interest on developing a new tool for miRNA target identification *in vivo*.

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Andrea Ventura, for providing me with the training and support essential for the completion of this work. Besides teaching me experimental skills, he also guided me into bioinformatics. The computational skills I have learned from him are critical for the development of my thesis project.

I would like to thank Dr. Carla P. Concepcion, who developed the initial HEAP protocol, and designed and generated the Halo-Ago2 knock-in mice. I would like to thank Drs. Christina Leslie, Yuri Pritykin, Yuheng Lu for sharing with me the unpublished CLIPanalyze algorithm and for the collaboration on this project.

I would like to thank Drs. Gaspare La Rocca and Minsi Zhang for sharing with me their unpublished RNA-seq datasets and for the insightful scientific discussions we had over the past few years. I would like to thank Paul Ogradowski and Chiara Mastroleo for their assistance in experiments.

I would like to thank my collaborators, Drs. Kevin M. Haigis, Olesja Popow and Joao A. Paulo for their help in the proteomic analysis. I would like to thank Dr. Ryan Schreiner for the help in live cell imaging.

I would like to thank my thesis committee members, Drs. Charles Rudin and Ross Levine for offering me guidance and support during the development of my thesis project. In particular, I would like to thank Dr. Ross Levine for his suggestion on combining the *in vivo* miRNA target identification with the various mouse models of human cancers generated in our laboratory, which became an integral part of this thesis.

I would like to thank our graduate school, especially Drs. Kenneth Mariani and Michael H. Overholtzer, Linda D. Burnley and David L. McDonagh, for the support and care that I have been given during the past years.

TABLE OF CONTENTS

LIST OF FIGURES	x
LIST OF TABLESxiii
LIST OF ABBREVIATIONSxiv
CHAPTER 1: INTRODUCTION	1
Overview of microRNAs	1
miRNA biogenesis	2
miRNA target recognition in mammals	5
Argonaute proteins	8
Mechanism(s) of miRNA-mediated gene repression	11
The biological functions of miRNAs	14
miRNA target prediction by computational algorithms	16
Experimental identification of miRNA targets	18
What Ago CLIP experiments teach us?	23
Limitations of CLIP-based methods	24
CHAPTER 2: A NOVEL EXPERIMENTAL APPROACH TO MAP MIRNA-RNA INTERACTIONS	26
Introduction	26
A Halo-Ago2 fusion protein enables antibody-free purification of miRNA tar- gets	27
Characterization of miRNA binding sites in mESCs	33
Reproducibility, sensitivity and detection limit of the HEAP method	36
Gene repression associated with HEAP-identified miRNA binding sites	39
Measuring translation inhibition associated with HEAP targets	42
Comparison between HEAP and the conventional CLIP method	45

Discussion	47
Method	50
CHAPTER 3: <i>IN VIVO</i> IDENTIFICATION OF MIRNA-RNA INTERACTIONS USING THE HALO-AGO2 MOUSE STRAIN	64
Introduction	64
Generation of the Halo-Ago2 knock-in mice	65
Characterization of the Halo-Ago2 mice	66
Characterization of the Halo-Ago2 fusion protein	70
Generation of HEAP libraries in vivo	76
Discussion	78
Method	81
CHAPTER 4: MIRNA TARGET IDENTIFICATION IN DEVELOPING EMBRYOS AND AUTOCHTHONOUS CANCER MODELS	86
Introduction	86
Characterization of miR-17~92 targets in developing embryos	88
Characterization of miRNA targets in Bcan-Ntrk1 gliomas and cortices	92
Characterization of miRNA targets in lung adenocarcinomas	98
Discussion	103
Method	106
CHAPTER 5: CONCLUSION AND PERSPECTIVE	112
Conclusions.	112
Future directions	113
APPENDIX I: Step wise protocol to generate HEAP and input control libraries	117
APPENDIX II: Computational pipeline used to process HEAP and input control libraries.	128
APPENDIX III: Oligonucleotides	134

APPENDIX IV: Example peaks in HEAP libraries	137
APPENDIX V: Generation of the <i>Lefty2</i> mutant mice	139
BIBLIOGRAPHY	144

LIST OF FIGURES

Figure 1.1: The canonical miRNA biogenesis pathway	3
Figure 1.2: Major target site types of mammalian miRNAs	6
Figure 1.3: Crystal structure of human AGO2	9
Figure 1.4: Summary of the major Ago CLIP variants	20
Figure 2.1: Subcellular localization, expression and endonucleolytic activity of the Halo-Ago2 fusion protein.	28
Figure 2.2: Generation of the Halo-Ago2 conditional knock-in allele	29
Figure 2.3: Schematic diagram of the HEAP method	31
Figure 2.4: HEAP libraries generated from mESCs	32
Figure 2.5: Genomic distribution of Halo-Ago2 binding sites	34
Figure 2.6: Enrichment of miRNA seed matches in HEAP 3'UTR binding sites .	35
Figure 2.7: Reproducibility of the HEAP method.	36
Figure 2.8: Sensitivity of the HEAP method to binding site disruption	37
Figure 2.9: HEAP libraries generated from decreasing numbers of mESCs. . .	38
Figure 2.10: Gene repression associated with Halo-Ago2 binding sites.	40
Figure 2.11: Correlation between miRNA abundance and gene repression capability	41
Figure 2.12: Benchmarking peak ranking parameters	42
Figure 2.13: Generation and characterization of the Dicer1 knockout mESCs .	43
Figure 2.14: Correlation between protein and mRNA expression in Dicer1 knock-out mESCs	44
Figure 2.15: Comparison between HEAP and iCLIP.	46
Figure 2.16: Comparison between miRNA binding sites identified by HEAP and TargetScan	48

Figure 3.1: Genotyping and protein expressed from the Halo-Ago2 knock-in allele	66
Figure 3.2: miR-451 isomiR shift in Halo-Ago2 expressing fetal livers.	68
Figure 3.3: Aberrant erythropoiesis in Halo-Ago2 knock-in embryos	69
Figure 3.4: Incorporation of Halo-Ago2 into miRISC	71
Figure 3.5: Efficient miRNA-mediated luciferase reporter repression in Halo-Ago2 expressing MEFs.	72
Figure 3.6: Efficient gene silencing and target slicing across a wide range of tar- get expression levels in Halo-Ago2 expressing MEFs	74
Figure 3.7: Gene expression profiling of Halo-Ago2 expressing MEFs	75
Figure 3.8: Comparison between HEAP and CLEAR-CLIP	77
Figure 3.9: Reduced <i>Ago2</i> mRNA expression from the Halo-Ago2 knock-in allele	78
Figure 4.1: miRNA target identification in mid-gestation embryos	87
Figure 4.2: Dose-dependent loss of miR-17~92 targets in miR-17~92 knockout embryos	88
Figure 4.3: Representative miR-17~92-dependent binding sites	90
Figure 4.4: Preferential target mRNA repression associated with HEAP binding sites	91
Figure 4.5: Regulation of a lncRNA by miR-92-3p	93
Figure 4.6: Differential expression of miRNAs between gliomas and normal corti- ces	94
Figure 4.7: Distinct miRNA targeting landscapes between gliomas and normal cortices.	95
Figure 4.8: Enrichment of miRNA seed matches correlates with miRNA abun- dance and causes differential repression of corresponding targets	97

Figure 4.9: Enrichment of miR-219a-5p binding sites in gliomas	98
Figure 4.10: miRNA target identification in lung adenocarcinomas and normal lungs	99
Figure 4.11: Similar miRNA targeting landscapes between the two NSCLC mod- els	101
Figure 4.12: Preferential gene repression associated with HEAP 3'UTR binding sites in KP cells	102
Figure 5.1: Live cell imaging of the Halo-Ago2 fusion protein	115
Figure 6.1: DNA sequencing chromatograms of <i>Lefty2</i> locus surrounding the miR-291-3p binding site	140

LIST OF TABLES

Table 3.1: Absolute numbers and frequencies of genotypes obtained from heterozygous intercrosses of *Ago2*^{Halo-LSL/+} or *Ago2*^{Halo/+} mice. 67

Table 6.1: *Lefty2* locus status of mice obtained from CRISPR-mediated gene targeting 141

LIST OF ABBREVIATIONS

miRNA: microRNA

Ago: Argonaute

HEAP: Halo-enhanced Ago2 Pulldown

miRISC: miRNA-induced silencing complex

3'UTR: 3'-untranslated region

5'UTR: 5'-untranslated region

CDS: coding sequence

LncRNA: long non-coding RNA

dsRNA: double-stranded RNA

ssRNA: single-stranded RNA

RNase III: Ribonuclease III

DGCR8: DiGeorge syndrome critical region gene 8

PAZ: Piwi-Argonaute-Zwille

PIWI: P-element induced wimpy testis

RBP: RNA binding protein

RNAi: RNA interference

B-CLL: B cell chronic lymphocytic leukemias

CLIP: crosslinking immunoprecipitation

HITS-CLIP: high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation

PAR-CLIP: photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation

iCLIP: individual-nucleotide resolution crosslinking immunoprecipitation

CLASH: crosslinking, ligation, and sequencing of hybrids

eCLIP: enhanced crosslinking immunoprecipitation

CLEAR-CLIP: covalent ligation of endogenous Argonaute-bound RNAs cross-linking immunoprecipitation

RIP: RNA immunoprecipitation

4SU: 4-thiouridine

MEF: mouse embryonic fibroblast

mESC: mouse embryonic stem cell

IRES: internal ribosomal entry site

PAGE (1): polyacrylamide gel electrophoresis

PAGE (2): parametric analysis of gene set enrichment

IDR: Irreproducible discovery rate

PhastCons: Phylogenetic analysis with space/time models conservation

CRISPR: clustered regularly interspaced short palindromic repeats

PAM: protospacer adjacent motif

sgRNA: single guide RNA

CDF: cumulative distribution function

TEV: tobacco etch virus

OL: oligodendrocyte

OPC: oligodendrocyte progenitor cell

NSCLC: non-small cell lung cancer

EA: Eml4-Alk-driven lung adenocarcinoma

KP: Kras^{G12D}; p53-loss lung adenocarcinoma

CHAPTER 1: INTRODUCTION

Overview of microRNAs

MicroRNAs (miRNAs) are ~22 nucleotide (nt) long, evolutionarily conserved, small non-coding RNAs (Bartel, 2004, 2009, 2018). Discovered in nematodes (Lee et al., 1993; Reinhart et al., 2000), this class of non-coding RNAs was later identified in a variety of eukaryotic lineages, including plants (Jones-Rhoades et al., 2006), invertebrates (Brennecke et al., 2003) and vertebrates (Lagos-Quintana et al., 2001).

To date, there are 2,656 and 1,978 annotated mature miRNAs in the human and mouse genomes, respectively (Kozomara et al., 2019). Non-coding RNAs and introns of protein coding genes are the major sources of miRNAs in cells (Okamura et al., 2007; Ruby et al., 2007). Through a series of ribonuclease (RNase) III cleavage events in cells, the precursor molecules are converted into mature miRNAs carrying 5'-phosphate and 3'-hydroxyl ends. Mature miRNAs associate with RNA binding proteins (RBPs) belonging to the Argonaute (Ago) family and function as negative regulators of gene expression at the post-transcription level. An essential step to fulfill their repressive role is the formation of miRNA-induced silencing complex (miRISC). miRNAs guide the miRISC to their targets by recognizing, via Watson-Crick pairing, cognate binding sites that are most commonly—but not exclusively—situated in the 3'-untranslated regions (3'UTR) of messenger RNAs (mRNAs). The miRISC contains enzymes that are necessary for inducing mRNA destabilization and/or translation inhibition, thus leading to reduction of the gene product (Fabian and Sonenberg, 2012).

The key determinant of miRNA targeting is the 5' region of the miRNA (also known as the seed), which pairs with complementary sequences on the targets. The miRNA seed is only 6 nt in length, making hundreds of transcripts potentially eligible for binding to a single miRNA. Given their high abundance and wide presence in almost every cell type, miRNAs are predicted to regulate around half of the protein coding genes (Friedman et al., 2009) and are involved in a wide range of cellular processes.

It is important to point out that the repressive effect caused by a single miRNA-target interaction is usually modest—resulting in less than two-fold changes in mRNA abundance. Nevertheless, such fine-tuning activity can be consequential to the delicate developmental programs in multicellular organisms. miRNA regulation modulates gene expression networks and promotes the robustness of cellular transcription programs (Ebert and Sharp, 2012). It shapes cell identity and plays vital roles in development and homeostasis. Its dysfunction has also been implicated in disease formation.

miRNA biogenesis

The majority of miRNAs are generated from primary-miRNAs (pri-miRNAs) transcribed from non-coding RNA genes by RNA polymerase II (Lee et al., 2002). Pri-miRNA can fold into a hairpin structure, which is processed by the microprocessor complex consisting of one molecule of Drosha and two molecules of DGCR8 (Denli et al., 2004; Han et al., 2004; Lee et al., 2003) (**Figure 1.1a**). Drosha belongs to the RNase III family and is a double-stranded RNA (dsRNA)-specific endonuclease (Court et al., 2013). Drosha binds to the lower stem of pri-miRNA and recognizes the UG motif present at the basal ssRNA-dsRNA junction (Auyeung et al., 2013; Nguyen et al., 2015). It acts as a “ruler” and cleaves at 11-bp from the

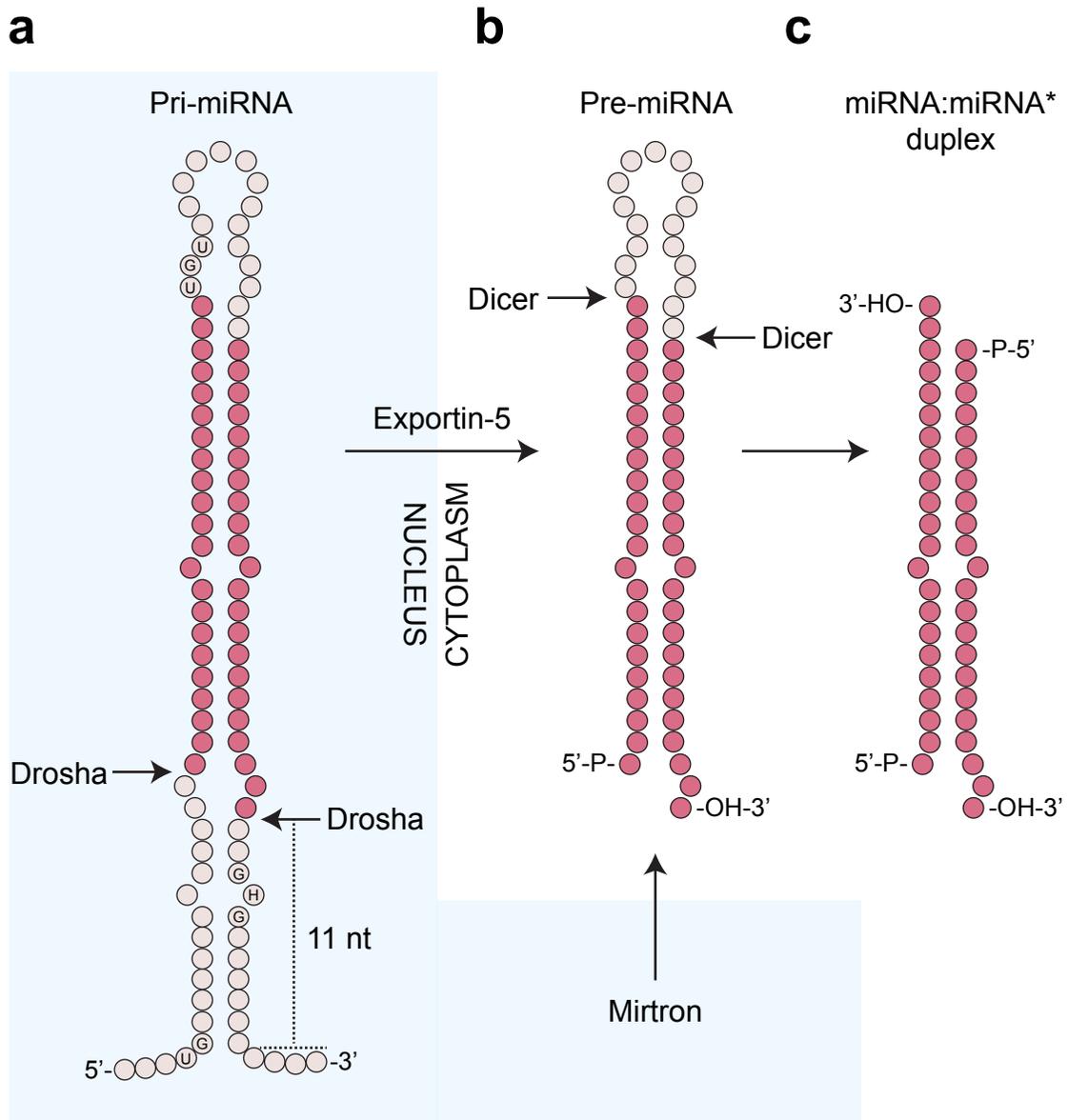


Figure 1.1: The canonical miRNA biogenesis pathway

a) Schematic of a pri-miRNA. Key features essential for pri-miRNA processing are annotated. In the nucleus, pri-miRNA is cleaved by Drosha in the microprocessor complex at positions marked by arrows, which produces a ~60-nt stem-loop RNA (pre-miRNA) as shown in **b)**. The pre-miRNA is exported into the cytoplasm by Exportin-5. **b)** Schematic of a pre-miRNA. Dicer cleaves the pre-miRNA close to its apical junction, which generates a miRNA:miRNA* duplex as illustrated in **c)**. **c)** A miRNA:miRNA* duplex (~22-nt in length) with 2-nt 3' overhangs.

UG motif, leaving a 2-bp offset. A newly identified mismatched GHG (mGHG) motif in the lower stem of pri-miRNA is recognized by the dsRNA binding domain of Dro-

sha and promotes accurate cleavage (Fang and Bartel, 2015; Kwon et al., 2019). DGCR8 is an RNA binding protein encoded by DiGeorge syndrome critical region gene 8. DGCR8 dimerizes and binds to the apical junction and the upper stem of the hairpin (by recognizing the UGU motif) to ensure efficient and precise pri-miRNA processing (Auyeung et al., 2013; Han et al., 2006; Nguyen et al., 2015). The cleavage leads to the release of a ~60-nt stem-loop RNA, known as the precursor miRNA (pre-miRNA) (Lee et al., 2002) (**Figure 1.1b**).

A subset of the pre-miRNAs can be directly generated from spliced introns of mRNAs, bypassing the need of the microprocessor complex. These introns/pri-miRNAs are known as “mirtrons” (Ladewig et al.; Okamura et al., 2007; Ruby et al., 2007). The conversion of pre-miRNAs from mirtrons relies on lariat debranching enzymes (Padgett et al., 1984; Ruskin and Green, 1985).

Pre-miRNAs are then exported by Exportin-5 into the cytoplasm in a Ran-GTP dependent manner (Lund et al., 2004; Yi et al., 2003), where they are further processed by a complex comprised of the endonuclease Dicer (Grishok et al., 2001; Hutvagner et al., 2001; Knight and Bass, 2001) and its binding partner, TRBP (Chendrimada et al., 2005; Gregory et al., 2005; Wilson et al., 2015) (**Figure 1.1b**). Similar to Drosha, Dicer is a member of the RNase III protein family. It cleaves the pre-miRNA close to the apical junction of the stem-loop, generating a miRNA:miRNA* duplex with 2-nt 3' overhangs (Zhang et al., 2004) (**Figure 1.1c**). Mechanistically, TRBP interacts with Dicer and enhances cleavage accuracy (Fukunaga et al., 2012; Kim et al., 2014; Wilson et al., 2015).

The miRNA:miRNA* duplex is loaded onto Ago proteins—the core effector proteins in the miRNA pathway—by the Dicer-TRBP complex (Chendrimada et al., 2005; Gregory et al., 2005). While the guide RNA (“miRNA”) is incorporated into the Ago proteins, the passenger strand (“miRNA*”) is discarded and degraded.

The unwinding and strand selection is dependent on the thermodynamics of the duplex, especially its 5' nucleotide identity (Gu et al., 2011; Khvorova et al., 2003; Schwarz et al., 2003; Tomari et al., 2004). The strand with the less stable 5' end is preferentially loaded onto Ago proteins. TRBP has been suggested to also influence strand selection, based on the observation that mutating the TRBP-interacting domain in Dicer causes aberrant guide strand selection (Wilson et al., 2015). However, it is unclear whether the activity of TRBP in regulating strand selection is independent from its role in promoting cleavage fidelity, because, in the absence of TRBP, the 5' identity of the 3p strand may be changed due to imprecise cleavage, which can lead to altered strand preference (Kim et al., 2014).

While the majority of pre-miRNAs are subjected to Dicer cleavage to produce mature miRNAs, pre-miR-451 maturation is Dicer-independent. Pre-miR-451 can fold into a short hairpin with perfect complementarity in its stem region, which becomes a good cleavage substrate for Ago2. After Ago2 cleavage, the miR-451 precursor remains associated with Ago2 and further matures by undergoing 3' end shortening (Cheloufi et al., 2010; Cifuentes et al., 2010; Yang et al., 2010).

miRNA target recognition in mammals

The initial characterization of miRNAs in *C. elegans* revealed that post-transcriptional regulation of the target mRNA is mediated by the direct interaction between the miRNA and mRNA molecules and that the 3'UTR of mRNA is the preferred target of mature miRNA (Ha et al., 1996; Lee et al., 1993; Wightman et al., 1993). Further studies revealed that targeting specificity is mainly determined by position 2-7 of the miRNA, also known as the miRNA "seed" (Lai, 2002; Lewis et al., 2003; Wightman et al., 1993). In mammals, having an exact match to the miRNA seed (6mer) is sufficient for miRNA binding. However, 6mer sites only margin-

ally contribute to target repression. Harboring an additional “A” on the mRNA opposite position 1 of the miRNA (7mer-A1) increases Ago binding affinity and target repression efficacy (Baek et al., 2008; Grimson et al., 2007; Nielsen et al., 2007; Schirle et al., 2014). A Watson-Crick match to position 2-8 of the miRNA (7mer-m8) is associated with even stronger repressive effect in comparison to a 7mer-A1 site

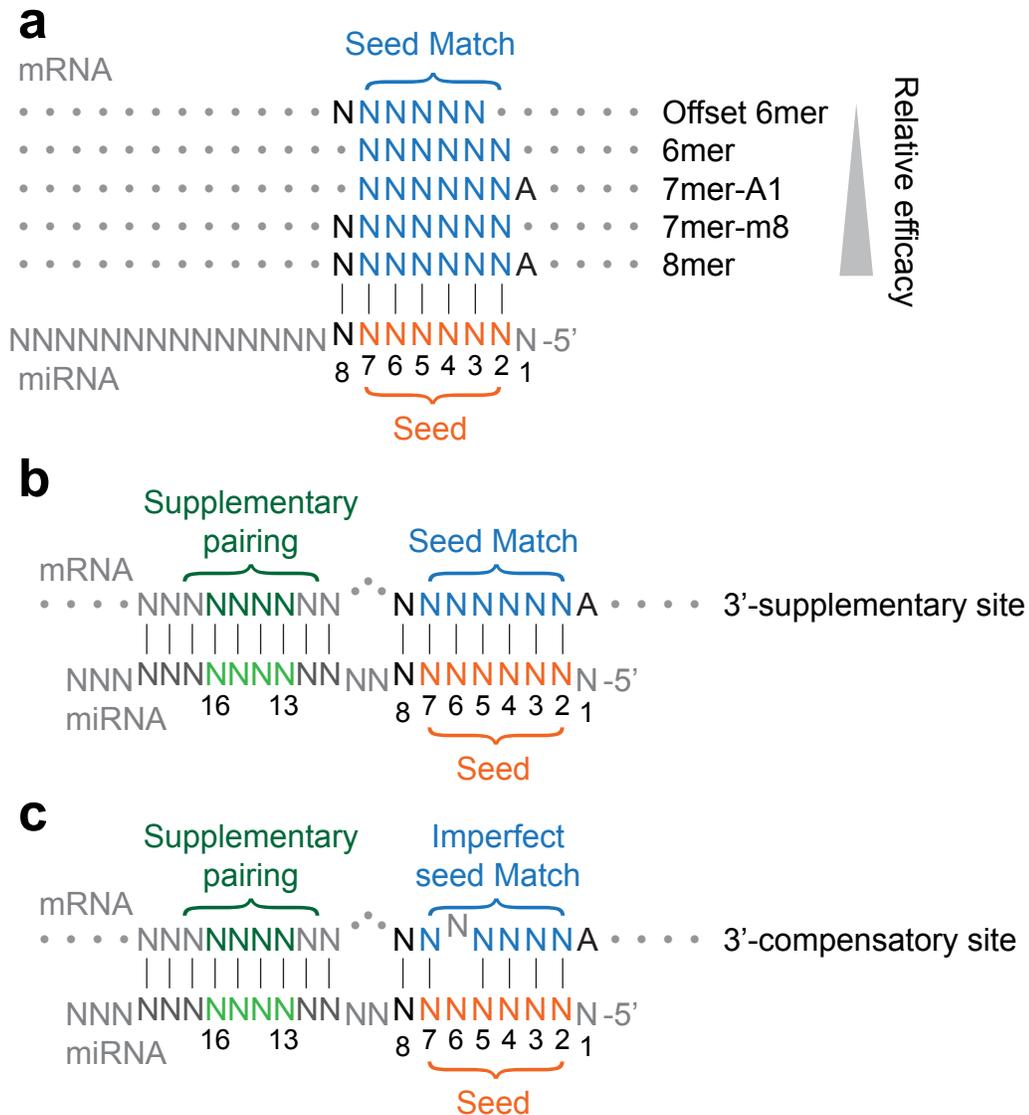


Figure 1.2: Major target site types of mammalian miRNAs

a) Canonical sites. b) Canonical site with supplementary pairing centered on position 13-16 of the miRNA. c) Non-canonical sites with 3' compensatory pairing. (Adapted from Bartel, 2018)

(Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005). Finally, sites with a Watson-Crick match to position 2-8 and an “A” opposite position 1 of the miRNA (8mer) experience the strongest repression (Lewis et al., 2005) (**Figure 1.2a**).

Structure analysis of the Ago2-miRNA-target complex suggests that additional pairing to position 9 and 10 of the miRNA does not further increase binding affinity (Schirle et al., 2014). However, additional base pairing with 3' of the miRNA, usually centered on position 13-16, moderately enhances site affinity and efficacy (3'-supplementary site) (Grimson et al., 2007; Salomon et al., 2015; Schirle et al., 2014; Wee et al., 2012) (**Figure 1.2b**). The site types mentioned above along with another marginal site type – offset 6mer (Friedman et al., 2009) – are collectively called canonical miRNA binding sites and account for the majority of miRNA-mediated gene repression detected in cells (Agarwal et al., 2015).

Biochemical approaches designed to identify miRNA binding sites transcriptome-wide (discussed in detail later in this chapter) have suggested that approximately half of the miRNA binding sites detected in cells do not have contiguous Watson-Crick pairs to the seed regions (Chi et al., 2009; Grosswendt et al., 2014; Hafner et al., 2012; Helwak et al., 2013; Loeb et al., 2012). These sites usually harbor additional Watson-Crick pairings in the 3' region of the miRNAs to compensate the lack of seed pairing and are called non-canonical sites (**Figure 1.2c**). Non-canonical sites only marginally contribute to target repression (Agarwal et al., 2015).

Because of the key role played by the seed sequence in determining target specificity, miRNAs with identical seed sequences, are grouped into the same family. A total of 1,558 miRNA seed families have been annotated in mammals, with 111 of them conserved across vertebrates (Agarwal et al., 2015).

Characterization of the targeting rules leads to several important predictions. 1) Members of the same seed family target similar mRNAs. 2) One miRNA

gene can theoretically regulate hundreds of mRNAs, as motifs complementary to the short 6-nt seed sequence can be easily encountered in the mammalian transcripts (Lim et al., 2005). 3) A single 3'UTR can be simultaneously regulated by multiple miRNAs, belonging to the same or different seed families. Computational studies have validated that more than half of the genes in the genome contain at least one conserved miRNA binding site and are potentially under miRNA control (Friedman et al., 2009).

Experimental data have suggested that miRNAs work in a highly redundant fashion. Multiple miRNAs synergistically control the same biological process, as deletion of a single miRNA gene often fails to result in a clear phenotype in model organisms. This functional cooperation has even been observed for miRNAs coming from different seed families. For example, miR-17 and miR-18 cooperatively regulate the transition from Pro- to Pre-B cells in the spleen of adult mice (Han et al., 2015). At pathway level, the functional redundancy ensures the safety of the regulation, making it less susceptible to genetic or environmental perturbations. From a systems biology point of view, the intricate miRNA regulation network, consisting of numerous “fail-safe” regulation units, promotes the robustness of the transcriptome and proteome.

Argonaute proteins

Members of the Argonaute (Ago) protein family belong to the PIWI protein superfamily and are deeply conserved across species. They are defined by the characteristic PAZ (Piwi-Argonaute-Zwille) and PIWI domains. Having high affinity for 5'-phosphorylated oligoribonucleotides, Ago proteins are the key components of the small RNA-mediated gene-silencing pathways (Farazi et al., 2008; Swarts et al., 2014). All Ago proteins contain 4 domains—the N(N-terminal), PAZ, MID (mid-

dle) and PIWI domains—connected by 2 linkers (L1 and L2) (**Figure 1.3**). Both MID and PAZ are RNA binding domains (RBDs). The MID domain anchors the 5' of mature miRNA by recognizing its 5'-phosphate. The PAZ domain interacts with the 3' end of the miRNA, especially the 2-nt 3' overhang in the miRNA:miRNA* duplex. The PIWI domain, highly similar to RNase H, retains catalytic activity in a subset of Ago members (Schirle and MacRae, 2012). Detailed structural analysis has revealed that Ago protein undergoes conformational changes to facilitate recognition and interaction between miRNA and its target. These studies also indicate that Watson-Crick pairing between the miRNA and target is, at least initially, restricted to the seed region (position 2-8) (Schirle et al., 2014), rationalizing the requirement of a seed match for miRNA targeting.

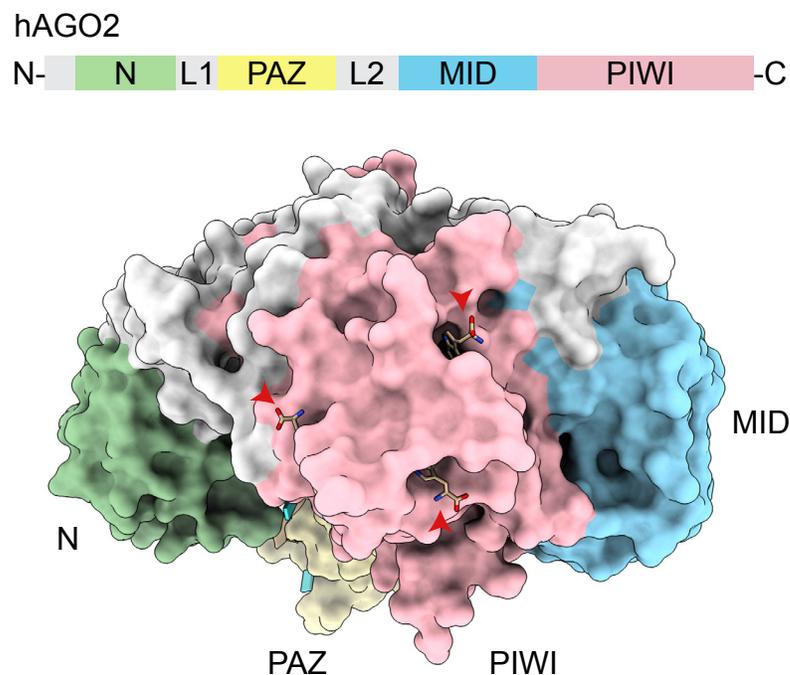


Figure 1.3: Crystal structure of human AGO2

Surface view of the human AGO2 structure (adapted from PDB: 6cbd). Red arrows: tryptophans inserted into tryptophan binding pockets identified on the surface of AGO2.

There are four Ago proteins (Ago1-4, encoded by *Eif2c1-4*) in the mouse genome. Ago1-4 are expressed across tissues and cell types. They bind to similar miRNAs, showing similar target-specificity when co-expressed in cells (Hafner et al., 2012). However, only Ago2 is essential for survival in mice. Ago2 knockout mouse embryos display developmental defects during gastrulation and die around embryonic day 10.5 (E10.5) (Liu et al., 2004). In contrast, mice double knockout for Ago1 and Ago3 are born at expected mendelian frequency with increased susceptibility to viral infection (Van Stry et al., 2012). A recent report suggested that while Ago4 is also dispensable for survival, it is involved in antiviral defense (Adil-iaghdam et al., 2020).

What distinguishes Ago2 from other members in the Ago protein family is the endonucleolytic activity inherited from its evolutionary ancestors. It is the three conserved metal-chelating residues – aspartate, aspartate and histidine (DDH) – in the PIWI domain that confer the slicing activity to Ago2 (Liu et al., 2004). Ago2 slices target RNA opposite position 10-11 of the guide RNA when the guide forms perfect complementarity to its target (Liu et al., 2004; Meister et al., 2004). This is also the mechanism underlying RNA interference (RNAi). Although this ancient defense mechanism against foreign dsRNA has been replaced by innate immune response in mammalian cells, endogenous RNAi can be detected in mouse oocytes and is essential for female fertility (Flemer et al., 2013; Wianny and Zernicka-Goetz, 2000). Ago2 has also been reported to direct the cleavage of *HOXB8* mRNA by miR-196 in mouse embryos (Yekta et al., 2004).

The embryonic lethal phenotype of Ago2 knockout mice can be in part explained by the loss of Ago2 catalytic activity. Mice mutant for the first aspartate residue in the DDH motif (*Ago2^{ADH}*) display erythroid differentiation defects and die before weaning age. The compromised viability in *Ago2^{ADH}* homozygous mice

is partially attributed to the lack of mature miR-451 in fetal livers, which plays essential roles in erythropoiesis and whose maturation is dependent on Ago2 catalysis (Cheloufi et al., 2010; Patrick et al., 2010). More recently, deregulation of another miRNA—miR-486-5p—in the absence of Ago2 catalytic activity has also been reported to contribute to the phenotype. While maturation of miR-486-5p is Dicer-dependent, it is sequestered in the miRNA:miRNA* duplex until the 3p strand is cleaved by Ago2. Mice knockout for miR-486 also experience erythropoietic defects (Jee et al., 2018). However, additional mechanisms are likely to contribute to the perinatal lethality phenotype observed in *Ago2^{ADH}* mice, as neither single deletion of miR-451 nor double deletion of miR-451 and miR-486 result in embryonic lethality (Jee et al., 2018; Patrick et al., 2010; Rasmussen et al., 2010; Yu et al., 2010).

Finally, the essentiality of Ago2 may result from spatial and temporal control of Ago proteins. A single-cell RNA profiling of early human and mouse embryos reveals that only Ago2 is expressed immediately after fertilization. Downregulation of Ago2 at later stages is accompanied by a slight upregulation of Ago1 (Boroviak et al., 2018; Müller et al., 2020). Losing Ago2 may not be tolerated by the embryo at the very early stages. The precise control of Ago2 expression during early development and cooperation among the Ago family members remain to be fully characterized.

Mechanism(s) of miRNA-mediated gene repression

In animals, Ago-miRNA complex represses the expression of its targets through the formation of a large ribonucleoprotein complex known as the miRNA-induced-silencing complex, or miRISC. The miRISC has been shown to act by repressing translation (Humphreys et al., 2005; Mathonnet et al., 2007; Olsen

and Ambros, 1999; Petersen et al., 2006; Pillai et al., 2005) and/or inducing mRNA decay via deadenylation and decapping (Behm-Ansmant et al., 2006; Chen et al., 2009; Giraldez et al., 2006; Rehwinkel et al., 2005; Wu et al., 2006).

miRNA targets were initially shown to localize to processing bodies (P-bodies) in an Ago-miRNA-dependent manner (Liu et al., 2005b; Sen and Blau, 2005). P-bodies are subcellular structures that serve as sites for mRNA degradation, enriching for untranslated mRNAs and enzymes involved in mRNA destabilization and translation inhibition (Sheth and Parker, 2003). The P-body component GW182 directly binds to Ago proteins and serves as a docking site for the other components of the miRISC that are responsible for target repression. As such, GW182 proteins are essential for miRNA-mediated gene silencing (Eulalio et al., 2008; Jakymiw et al., 2005; Liu et al., 2005a; Meister et al., 2005).

In mice and humans, GW182 are encoded by three *Tnrc6* genes (*Tnrc6a*, *Tnrc6b* and *Tnrc6c*) and characterized as 182-kDa proteins containing multiple glycine-tryptophan (GW) repeats (Baillat and Shiekhhattar, 2009; Eystathioy et al., 2002). The GW repeats are concentrated in a large unstructured region in the N terminus that mediates direct interaction with Ago proteins. The interaction between Ago and GW182 proteins is multivalent, with one GW182 molecule making contact with multiple Ago proteins simultaneously (Elkayam et al., 2017). Structural studies have determined three tryptophan-binding pockets on the surface of Ago proteins responsible for GW182 binding (Elkayam et al., 2017; Schirle and MacRae, 2012; Sheu-Gruttadauria and MacRae, 2018) (**Figure 1.3**). Accordingly, the key tryptophan residues on GW182 for Ago binding have also been mapped (Hauptmann et al., 2015; Pfaff et al., 2013; Sheu-Gruttadauria and MacRae, 2018).

GW182 protein acts as a scaffold in the miRISC, recruiting the CCR4-NOT and PAN2-PAN3 deadenylase complexes for target mRNA deadenylation (Braun

et al., 2011; Chekulaeva et al., 2011; Fabian et al., 2011; Fabian and Sonenberg, 2012; Kuzuoğlu-Öztürk et al., 2012). The C terminus of GW182 directly interacts with PABP, a poly(A)-binding protein involved in mRNA translation and metabolism (Fabian et al., 2009). A recent study shows that miRISC phase separates in cells and the induction of phase separation accelerates target mRNA deadenylation (Sheu-Gruttadauria and MacRae, 2018). The decapping enzymes DCP1 and DCP2 in P-bodies are also involved in miRNA-mediated mRNA degradation (Behm-Ansmant et al., 2006; Chen et al., 2009; Rehwinkel et al., 2005), but it is unclear whether they are in direct contact with GW182. Additionally, the interaction between an E3 ubiquitin ligase EDD and GW182 has been proven consequential for miRNA-mediated gene silencing in mouse embryonic stem cells (mESCs) (Su et al., 2011). The ubiquitin ligase activity of EDD is dispensable for the repressive effect. EDD likely interacts with the DEAD box helicase DDX6/RCK/p54 and promotes cap-dependent translation inhibition (Chu and Rana, 2006; Collier and Parker, 2005; Su et al., 2011).

The outcome of miRNA-mediated repression is likely dependent on the effector proteins recruited to the miRISC in a given context. In mammalian cells, mRNA destabilization explains the majority of target repression caused by miRNAs (Eichhorn et al., 2014; Guo et al., 2010). However, during maternal-to-zygote transition of zebrafish embryos, miR-430 is used to switch off protein translation on maternally expressed mRNAs before these transcripts get cleared by deadenylation processes (Bazzini et al., 2012; Giraldez et al., 2006).

The miRISC has been long assumed to be constitutively active in cells. Recent evidence, however, suggests that the activity of Ago proteins is dynamically regulated by a series of phosphorylation and dephosphorylation events (Golden et al., 2017) and the assembly of miRISC is contingent upon cellular states (La Roc-

ca et al., 2015). In adult mice under homeostatic condition, only a subset of tissues contain fully assembled miRISC (La Rocca et al., 2015), though the physiologic relevance of this observation, and its functional consequences, have not been experimentally addressed yet.

The biological functions of miRNAs

The biological roles of miRNAs in model organisms have been systematically examined using genetic loss-of-function approaches. As suggested by the seminal studies which led to the discovery of miRNAs, these small non-coding RNAs are involved in controlling developmental timing (*let-7*, *lin-4*), larval viability (*let-7*) and left/right asymmetry differentiation (*lisy-6*) of *C. elegans* (Johnston and Hobert, 2003; Lee et al., 1993; Reinhart et al., 2000). However, follow-up phenotypic characterizations of other miRNA mutants have suggested that the majority of miRNA families are dispensable for the development of *C. elegans* (Miska et al., 2007). Inactivation of *dcr-1*, a homolog of mammalian Dicer, in *C. elegans* produces heterochronic phenotypes similar to *lin-4* and *let-7* mutants (Grishok et al., 2001). It is tempting to speculate that most miRNAs in *C. elegans* are not functional under the experimental conditions or have largely overlapping roles.

In contrast to the miRNAs in *C. elegans*, in *Drosophila Melanogaster*, ~80% of the miRNA families are required for fly development and survival (Chen et al., 2014). In zebrafish, miRNAs are essential for morphogenesis, as embryos lacking maternal-zygote Dicer activity display abnormal gastrulation, impaired brain formation and somite defects (Giraldez et al., 2005).

In mouse, deletion of key enzymes in the miRNA biogenesis pathway leads embryonic lethality at very early stages (Bernstein et al., 2003; Liu et al., 2004). Mice knockout for many broadly conserved miRNA families display a wide spec-

trum of abnormalities (reviewed in (Bartel, 2018)). It is important to point out that despite the defects observed in these studies, deletion of a single miRNA gene often fails to cause a phenotype. Abnormalities are often observed only after ablating all members of the same miRNA family. Moreover, certain miRNAs may be dispensable for development under normal conditions but required upon stresses and challenges. Therefore, the biological role of a miRNA must be evaluated under physiological condition.

In human, the type 2 Feingold syndrome, a disease characterized by a series of skeletal abnormalities in patients, is caused by mutations affecting the MIR17 host gene (*MIR17HG*) (Feingold et al., 1997). Mice haploinsufficient for this miRNA cluster (*miR-17~92*) phenocopy the several key features of the patients (de Pontual et al., 2011). The *DGCR8* gene resides in a 3Mb region on chromosome 22 (Chr 22q11.2) that is deleted in the DiGeorge syndrome (Shiohama et al., 2003; Wilson et al., 1993). *DGCR8* loss may contribute to the pathogenesis of DiGeorge syndrome as haploinsufficiency of the *Dgcr8* gene in mice leads to abnormal miRNA biogenesis, as well as cognitive and behavior deficits (Stark et al., 2008).

De-regulation of the miRNA pathway also contributes to tumorigenesis (Di Leva et al., 2014). Heterozygous germline *DICER1* mutations are associated with a rare pediatric lung tumor – pleuropulmonary blastoma (Hill et al., 2009). In line with the observation made in human cancer, Dicer1 functions as a haploinsufficient tumor suppressor in mouse models of lung cancer and sarcoma (Kumar et al., 2007; Kumar et al., 2009).

Several miRNAs have been suggested to be tumor suppressive. For example, the locus harboring miR-15 and miR-16 on the long arm of chromosome 13 is lost in more than 50% of B cell chronic lymphocytic leukemias (B-CLL) (Calin et al., 2002). *BCL2*, overexpressed in B-CLL and an antiapoptotic gene, is likely

to be a target of miR-15/16. Escaping miR-15/16 repression maybe beneficial for B-CLL formation (Cimmino et al., 2005; Klein et al., 2010). Additionally, members of the miR-200 family work as suppressors for epithelial-to-mesenchymal transition (EMT) by targeting key factors, such as *Zeb1* and *Zeb2*, in the pathway (Gregory et al., 2008; Korpál et al., 2008; Park et al., 2008).

The first reported pro-tumorigenic miRNA is miR-155. miR-155 host gene, also known as *bic*, was initially identified as a locus susceptible to viral integration in avian lymphomas and was shown to cooperate with *c-myc* in B-cell lymphomagenesis (Clurman and Hayward, 1989; Tam et al., 1997; Tam et al., 2002). Up-regulation of miR-155 along with its host gene was later observed in human B-cell lymphomas (Eis et al., 2005). One of the most extensively characterized oncogenic miRNAs are those encoded by the miR-17~92 miRNA cluster, also known as “Oncomir-1”. miR-17~92 is comprised of six miRNA members (miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1 and miR-92a). Frequent amplification of the locus containing miR-17~92 in lymphomas provided the first hint for its oncogenic roles (Ota et al., 2004). Later, this miRNA cluster was found to be a transcription target of c-Myc (O’Donnell et al., 2005) and overexpressed in several MYC-driven cancers (Han et al., 2015; He et al., 2005). Functional dissection of this cluster suggested that the tumorigenic activity of miR-17~92 can be attributed to miR-19, as deleting miR-19 in the E μ -Myc lymphomas extends mouse overall survival and induces apoptosis (Han et al., 2015; Mu et al., 2009; Olive et al., 2009).

miRNA target prediction by computational algorithms

Key to understand the biological roles of a given miRNA is the identification of its physiologic targets. The obvious complementarity between miRNAs and a set of 3’UTRs under post-transcriptional repression in drosophila for the first time sug-

gested the possibility of *in silico* miRNA target prediction (Lai, 2002). Since then, a variety of algorithms have been proposed to computationally predict the targets of miRNAs (reviewed in (Bartel, 2009)) and their performance has greatly improved over the past decade, thanks to the accumulation of experimental evidence.

Arguably, the most widely accepted prediction algorithm is TargetScan (Agarwal et al., 2015; Friedman et al., 2009; Grimson et al., 2007; Lewis et al., 2003). In the first versions of Targetscan, targets of conserved miRNAs were predicted in 3'UTRs that are broadly conserved across species. Using targets predicted for shuffled miRNAs as background (false-positives), target sites were scored and ranked using the free energy of miRNA:target interactions (Lewis et al., 2003). Later, as genomes of more vertebrate species became available, a larger set of mRNAs with conserved complementarity to miRNA seeds were predicted. Interestingly, including a highly conserved adenosine ("A") opposite position 1 of the miRNA strongly enhanced the performance of the prediction (Lewis et al., 2005). The benefit of retaining the "A" opposite position 1 was not clear until the Ago2-miRNA-mRNA co-crystal structure was solved and the binding pocket in Ago2 for the adenosine was revealed (Schirle et al., 2014).

miRNA transfection experiments followed by gene expression profiling allowed the experimental validation of predicted binding sites. These efforts revealed that the efficacy of the predicted canonical miRNA binding sites, measured as preferential downregulation their targets, correlates with the number of nucleotides involved in the interactions, with 8mer sites being the most effective, followed by 7mer-m8, 7mer-A1 and 6mer sites. Furthermore, this also enabled miRNA targeting efficacy predictions by integrating the contribution of several context features, including the 3' pairing score, local AU content and distance from the nearest 3'UTR terminus (Grimson et al., 2007).

Finally, improvement in conservation analysis and growing number of datasets measuring cellular response to miRNA perturbations led to the development of more powerful prediction algorithms. To date, more than half of human transcripts are predicted to be under miRNA regulation (Agarwal et al., 2015; Friedman et al., 2009).

In addition to TargetScan, a series of prediction tools were generated around the same time, including PicTar (Lall et al., 2006), EMBL (Stark et al., 2005) and miRanda (Betel et al., 2008). These approaches, though different in the criteria used for predictions, are all filtering algorithms based on predefined features. Later on, another group of prediction tools were developed that used machine learning algorithms. These algorithms examine efficacy of all possible miRNA-target interactions using gene expression datasets in transfection experiments and build prediction models by integrating parameters characterizing miRNA-target duplex and context features. One example of such tools is miRSVR. miRSVR applied the support vector regression algorithm and built a scoring system that improved the ranking of canonical sites and identified functional non-canonical and non-conserved sites (Betel et al., 2010).

Experimental identification of miRNA targets

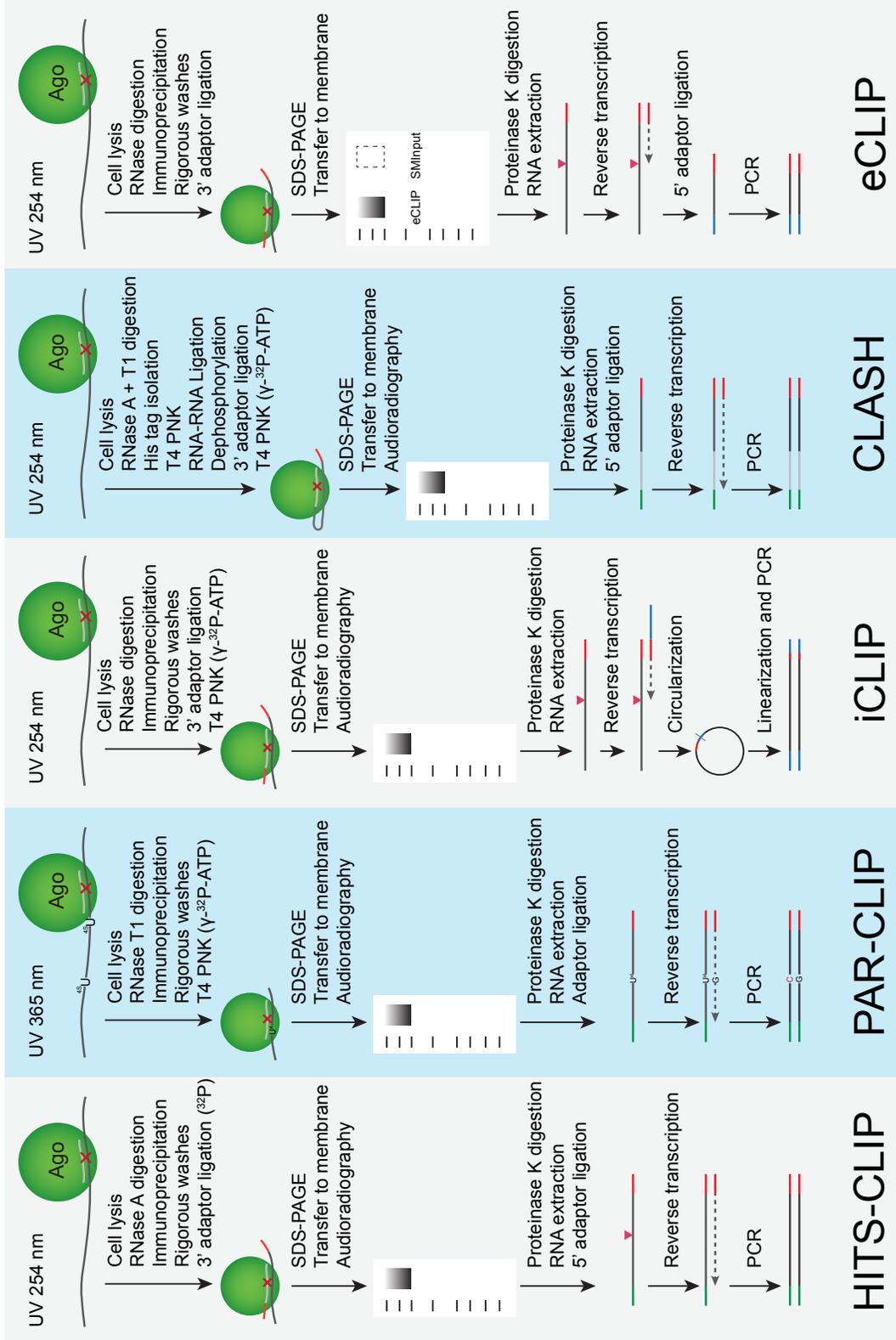
Despite the increasing sophistication of these computational approaches, the usefulness of their predictions is limited by the fact that computational algorithms do not take into consideration the actual abundance of the miRNA and its predicted targets in a given cellular context, or the presence of additional factors that can either impede or promote miRNA binding to a specific site (Kedde et al., 2010). To overcome these limitations, a number of experimental approaches aimed at directly identifying miRNA-mRNA interaction sites in cells and tissues

have been developed over the past decade (Chi et al., 2009; Hafner et al., 2012; Helwak et al., 2013; Konig et al., 2010; Van Nostrand et al., 2016).

Crosslinking immunoprecipitation (CLIP) was first described by Ule and colleagues in 2003 as a novel approach for identifying direct targets of the RBP Nova in mouse brain (Ule et al., 2003). In this approach, RNAs are first immobilized on RBPs in living cells by UV crosslinking. Following cell lysis, sequences flanking the core interaction motifs are trimmed off by ribonuclease. The crosslinked complexes are purified from the lysate using a series of biochemical approaches, including immunoprecipitation of the RBP, RNA radiolabeling, and SDS-polyacrylamide gel electrophoresis (SDS-PAGE) purification. The complexes are then transferred onto a nitrocellulose membrane and the radiolabeled RNAs are released from the complexes by Proteinase K treatment. Finally, isolated RNAs are reverse transcribed into cDNAs, which are amplified and sequenced. The initial CLIP experiment identified RNA targets enriching for Nova binding sites and implicated in synapse functions. The early success also paved the way for applying similar approach for miRNA target discovery *in situ* in live cells.

As high-throughput sequencing technologies became available, they were employed for RBP target discovery together with the CLIP method. The very first version of such technique was named high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation (HITS-CLIP) and was applied to isolate Ago2-bound miRNA targets (Chi et al., 2009) (**Figure 1.4**).

Due to the close proximity between Ago2, miRNA and target mRNA at the site of interaction, the components in this complex can be crosslinked using 254-nm UV-C. In their pioneering work, Chi and colleagues immunoprecipitated cross-linked Ago2-RNA complexes from freshly prepared cortex tissues of post-natal day 13 (P13) mice and purified both the miRNAs and target mRNAs. High-throughput



✕ Crosslinking site
 ▶ Crosslinked polypeptide

Figure 1.4: Summary of the major Ago CLIP variants

sequencing of target libraries identified a large number of binding sites preferentially in the 3'UTR of mRNAs. A fraction of the binding sites overlapped with predicted or experimentally validated sites for miR-124-3p (Chi et al., 2009). It is important to highlight that, at the reverse transcription step, the crosslinked residue presents a roadblock for reverse transcriptase, which often leads to truncation of the cDNA at the crosslinking site or skipping of the ribonucleotide. While the truncated cDNAs are lost during subsequent amplification and sequencing steps, the skipping events, which are later translated into deletion events in sequencing, help impute the precise interaction sites at single-nucleotide resolution (Moore et al., 2014).

Soon after the advent of the HITS-CLIP method, several variants were generated, addressing specific needs in the field. For example, Hafner and colleagues developed photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation (PAR-CLIP) to improve RNA recovery and increase the resolution of the method (**Figure 1.4**). In PAR-CLIP, cells are pre-incubated with the ribonucleoside analog 4-thiouridine (4SU), which replace uridines in RNAs and can be more efficiently crosslinked to RBPs by 365-nm UV at the same radiation energy. Another benefit offered by this strategy is thymidine (T)-to-cytidine (C) transitions occurring at the crosslinking sites in the reverse transcription step, which, again, provide single-nucleotide resolution information about the interaction sites (Hafner et al., 2010). In the first AGO PAR-CLIP experiment, an enrichment of miRNA seed matches in the target sequences was observed and T-to-C transitions were frequently found around putative seed matches, indicating purification of *bona fide* miRNA binding sites *in vitro*.

Another widely used CLIP method is individual-nucleotide resolution crosslinking immunoprecipitation (iCLIP) (Konig et al., 2010) (**Figure 1.4**). To retrieve information from truncated cDNAs produced at the reverse transcription step and

map precise crosslinking sites, an intramolecular circularization step is added to the workflow followed by linearization and PCR amplification. As self-circularization is more efficient than intermolecular ligation, the sensitivity of this method surpasses its precedents.

Despite the simultaneous purification of miRNA and targets in Ago CLIP experiments, assigning targets to individual miRNAs remains bioinformatically challenging as one has to indirectly infer the bound miRNA from the target sequence. A way to solve this problem is to modify the CLIP method to promote the intermolecular ligation between the miRNA and its target in the complex. Helwak and colleagues were the first to propose this strategy, which they named crosslinking, ligation, and sequencing of hybrids (CLASH) (Helwak et al., 2013) (**Figure 1.4**). Although chimeric reads only represented 2% of all sequenced reads, CLASH identified interactions between 399 miRNAs and close to 7,000 protein-coding genes. Compared to canonical sites, surprisingly, non-canonical binding sites were ~1.7-fold more frequently captured. These non-canonical sites only subtly, though significantly, contributed to target gene repression as validated by luciferase reporter assays. Two similar approaches: a variant of the *in vivo* PAR-CLIP (iPAR-CLIP) (Grosswendt et al., 2014) and covalent ligation of endogenous Argonaute-bound RNAs-CLIP (CLEAR-CLIP) (Moore et al., 2015) were later applied to capture miRNA-target hybrids in *C. elegans* and mouse cortex, respectively.

To date, the most widely used CLIP method is enhanced CLIP (eCLIP) (Van Nostrand et al., 2016) (**Figure 1.4**). The eCLIP method omits the radiolabeling step and replaces the circularization reaction used in the iCLIP with single-stranded DNA (ssDNA) adaptor ligation following reverse transcription. These modifications improve RNA recovery efficiency, reduce PCR duplication rate and shorten hands-on time. In addition to the optimized library preparation strategy, eCLIP introduc-

es a “SMInput” library, which is produced from adapter-ligated RNAs co-migrating with the immunoprecipitated complexes in the RNase-treated input lysates. The SMInput is used as background and greatly improves signal-to-noise and the reproducibility of eCLIP experiments. eCLIP has recently been employed to generate the binding site maps for 150 human RBPs including AGO2 in the ENCODE 3 project (Van Nostrand et al., 2020).

What Ago CLIP experiments teach us?

When applied to Ago proteins, the CLIP methods described in the previous section have nearly invariably shown that 3'UTRs are the preferred regions targeted by miRNAs. Motif discovery analyses have identified seed matches for highly expressed miRNAs in each context. And all previously characterized canonical site types (6mer, 7mer-A1, 7mer-m8 and 8mer) have been observed in CLIP datasets. While a certain fraction of these binding sites matches computationally predicted sites, a partial overlap between the experimentally determined and predicted binding sites suggested both false-positives and -negatives in the prediction results. When benchmarking these sites using context-matched gene expression datasets, CLIP-identified binding sites often outcompeted those given by prediction algorithms in conferring target repression (Hsin et al., 2018). This again emphasizes the need of determining miRNA binding sites in physiological contexts.

In addition to 3'UTR binding sites, a substantial fraction of binding sites has been observed in the coding exons (Hafner et al., 2010). These interactions sites are less extensively characterized compared to 3'UTR binding sites as high conservation status of coding sequences (CDS) prevents filtering of these miRNA binding sites using conventional target prediction algorithms. In contrast to 3'UTR sites that preferentially cause mRNA destabilization, CDS sites, as suggested by

computational and experimental studies, are likely to be more effective in inhibiting translation, though a detailed molecular mechanism is still unclear (Hausser et al., 2013; Zhang et al., 2018).

Limitations of CLIP-based methods

Despite the robustness of the various CLIP-based approaches in identifying biologically relevant miRNA binding sites, these methods still suffer from several limitations, which have prevented their wider use. First, although the technical workflow has been simplified in recent variants, such as in eCLIP, these methods are still quite technically challenging. They often require a large number of cells as input to compensate for material loss during the purification procedures. Second, the affinity of the antibody limits the purification stringency and extensive optimization experiments are required to determine the highest salt concentration that can be tolerated for a given antibody-epitope combination.

The use of 4SU in PAR-CLIP improves RNA-RBP crosslinking efficiency, however, it also biases towards identification of targets enriching for uridines. Additionally, precise mapping of interaction sites becomes challenging near U-rich regions. It is worth to point out that prolonged 4SU treatment slows down cell growth (Tani and Akimitsu, 2012) and 4SU may subtly alter cell physiology by affecting RNA turnover and dynamics. Finally, 4SU incorporation is only applicable to cells in culture and *C. elegans* (Jungkamp et al., 2011). Such method is not suitable for *in vivo* miRNA target discovery for most model organisms.

Capturing miRNA-target hybrids following ligation by CLASH or CLEAR-CLIP (Helwak et al., 2013; Moore et al., 2015) still presents a promising approach for dissecting functional cooperation and specialization among miRNA members. However, the sensitivity of current approaches is largely compromised by the low

chimeric rate (~1-2%).

Given the highly context-dependent nature of miRNA targeting, it is crucial to perform target identification experiments in physiologic contexts. Understanding the physiologic roles of miRNAs often requires efficient target isolations *in vivo* and sometimes from specific cell types. Although target purification from live tissues has been proven feasible, none of the current CLIP methods offers a convenient workflow for Ago2-RNA complex isolation *in vivo*, especially in a tissue- and cell-type-specific manner.

Therefore, an efficient, unbiased *in vivo* miRNA target purification strategy is in demand.

CHAPTER 2: A NOVEL EXPERIMENTAL APPROACH TO MAP MIRNA-RNA INTERACTIONS

Introduction

Although Ago2 CLIP methods have proven useful in identifying miRNA-mRNA interaction sites in cell lines (Chi et al., 2009; Hafner et al., 2010; Helwak et al., 2013; Konig et al., 2010; Van Nostrand et al., 2016), they have been less extensively used in *in vivo* settings mainly for two reasons. First, these methods are technically challenging, and their performance is substantially reduced as the quality and quantity of the starting material drop. Second, live tissues are comprised of a variety of cell types and these methods do not provide an efficient way to purify targets from specific cell types without preliminary cell enrichment. To circumvent the difficulties presented by the CLIP methods, we decided to design a new system that permits high-affinity Ago2 complex purification and provides flexibility for potential *in vivo* applications.

The purification stringency of traditional CLIP-based approaches is limited by the strength of the non-covalent interaction between the antibody and epitope at the experimental condition. Therefore, when searching for alternative strategies, we specifically focused on protein tags that mediate covalent interactions, which would theoretically be resistant to stringent purifications. The candidate protein tags include the SNAP-tag (Keppler et al., 2003), CLIP-tag (Gautier et al., 2008) and Halo-tag (Los et al., 2008). These protein tags are of comparable sizes and can form irreversible covalent bonds with their corresponding synthetic ligands. Conjugating the ligands to solid substrates, such as agarose beads, enables rapid

isolation of protein of interest fused to these synthetic tags.

Following a suggestion from our colleague Gregory Hannon, we chose the HaloTag, which is a modified version of the haloalkane dehalogenase encoded by the *DhaA* gene from *Rhodococcus rhodochrous*. Haloalkane dehalogenase removes halides from aliphatic hydrocarbons through forming an intermediate ester bond between an aspartate and the substrate. A mutation in the histidine His272 blocks the hydrolysis of the ester bond, resulting in irreversible attachment of the chemical functionalities (Encell et al., 2012; Los et al., 2008). Due to the absence of homologs to *DhaA*, cross-reactivity of HaloTag with endogenous substrate is negligible in mammalian cells. The many applications of the HaloTag technology are enabled by the commercial availability of a wide range of synthetic HaloTag ligands, generated by linking the chloroalkane substrate to different functional groups, including fluorescent dyes, solid surfaces, and reactive groups.

In this work we envisioned to use the HaloTag system to simplify the isolation of Ago2-containing complexes from cells and live tissues. Importantly, while our work was in progress, Gu and colleagues showed that the HaloTag together with UV crosslinking could be used to efficiently identify RNA targets of the RBP PTB (Gu et al., 2018).

A Halo-Ago2 fusion protein enables antibody-free purification of miRNA targets

Because of previous reports showing that Ago2 does not tolerate C-terminal tags, we fused the HaloTag to its N-terminus (**Figure 2.1a**). When expressed in *Ago2*^{-/-} mouse embryonic fibroblasts (MEFs) (O'Carroll et al., 2007), the Halo-Ago2 fusion protein localized largely to the cytoplasm, while the HaloTag alone displayed uniform localization to both the cytoplasm and the nucleus (**Figure 2.1b**). The

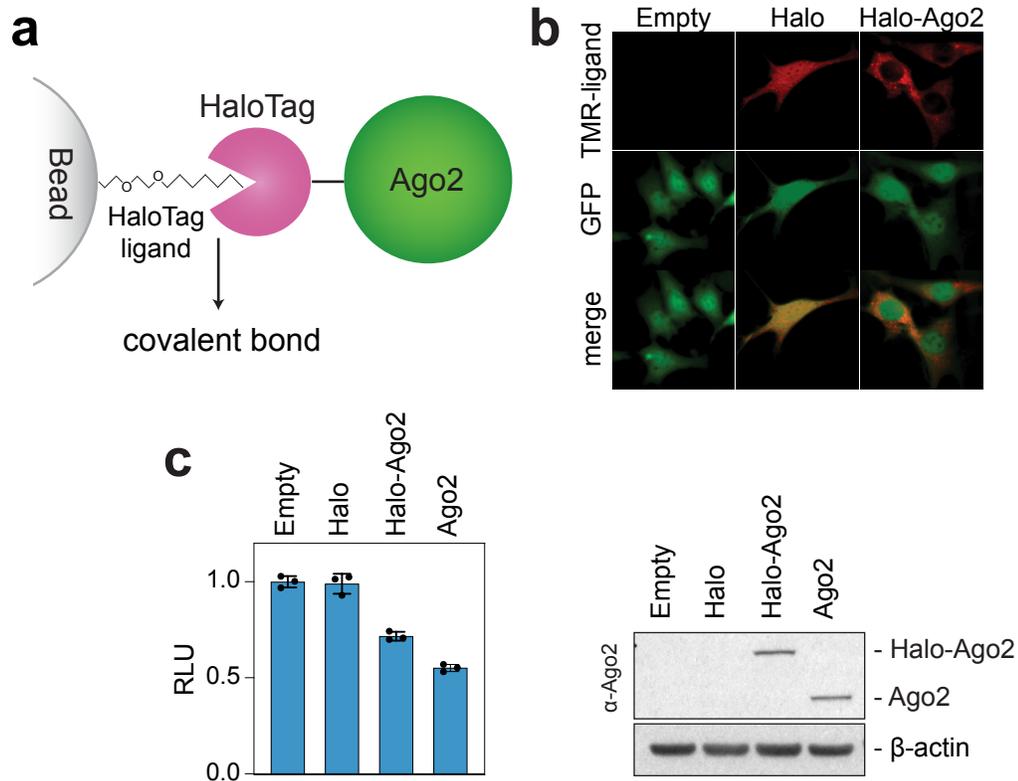


Figure 2.1: Subcellular localization, expression and endonucleolytic activity of the Halo-Ago2 fusion protein

a) Schematic of the Halo-Ago2 fusion protein covalently bound to a bead-conjugated HaloTag ligand. **b)** *Ago2*^{-/-} immortalized MEFs transduced with MSCV-PIG, MSCV-PIG-Halo or MSCV-PIG-Halo-Ago2 retroviruses were incubated with the HaloTag TMRDirect ligand and imaged (by Joana A. Vidigal). **c)** *Ago2*^{-/-} MEFs transduced with retroviral vectors encoding HaloTag alone, full length Ago2 or the Halo-Ago2 fusion protein were transiently transfected with reporter plasmids expressing Firefly and Renilla luciferase and a plasmid expressing an shRNA against the Firefly luciferase. The ratio between Firefly and Renilla luciferase activity was measured 48 hours after transfection (left panel). Whole-cell lysates from the same cells were probed with antibodies against Ago2 and β -actin (right panel). Error bars: Mean \pm SD.

generation of the Halo-Ago2 fusion construct and the imaging experiment were conducted by Joana A. Vidigal, a former postdoctoral fellow in the laboratory. To test whether the addition of the HaloTag interferes with the activity of Ago2, we transduced *Ago2*^{-/-} MEFs at low multiplicity of infection (MOI < 0.1) with recombinant retroviruses expressing either wild-type Ago2 or the Halo-Ago2 fusion protein.

Next, we transiently transfected these cells with reporter plasmids expressing the Renilla and Firefly luciferase and a plasmid expressing an shRNA against the Firefly luciferase and measured the ratio between Firefly and Renilla luciferase activity 48 hours later. These experiments showed that when expressed at comparable levels, the Halo-Ago2 construct was nearly as effective as wild-type Ago2 at rescuing RNA interference in *Ago2*^{-/-} MEFs (**Figure 2.1c**).

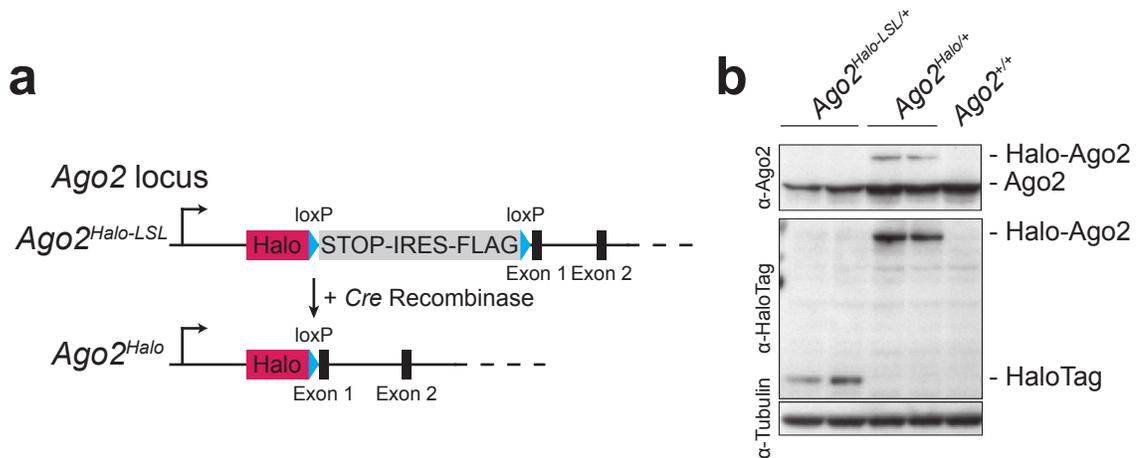


Figure 2.2: Generation of the Halo-Ago2 conditional knock-in allele

a) Schematic of the targeting strategy used to generate the Halo-Ago2 conditional knock-in allele. Halo: HaloTag; STOP: stop codon; IRES: internal ribosome entry site. **b)** Whole-cell lysates from mESCs with the indicated genotypes were probed with antibodies against Ago2, HaloTag and Tubulin. (The mESC targeting was designed and performed by Carla P. Concepcion and Joana A. Vidigal. The western blot was performed by Carla P. Concepcion)

To avoid artifacts due to ectopic expression of Halo-Ago2 and to enable the isolation of Ago2 complexes directly from murine tissues, we used homologous recombination to knock-in the HaloTag cassette into the endogenous *Ago2* locus in mESCs (**Figure 2.2a**). In this knock-in allele, the HaloTag is separated from the first coding exon of Ago2 by an in frame loxP-STOP-IRES-FLAG-loxP (LSL) cassette (*Ago2*^{Halo-LSL}). Cells harboring this allele express a bicistronic mRNA encoding

for two proteins: the HaloTag and a Flag-Ago2 fusion protein whose translation is initiated by an internal ribosomal entry site (IRES). Upon expression of the Cre recombinase, the LSL cassette is excised and the HaloTag is now brought in frame with the first coding exon of Ago2, thus resulting in expression of the Halo-Ago2 fusion protein (**Figure 2.2b**). The recombined allele expressing the Halo-Ago2 fusion will be hereafter referred to as *Ago2^{Halo}*. The mESC targeting was designed and performed by Joana A. Vidigal and Carla P. Concepcion, a former postdoctoral fellow and a former graduate student in the laboratory.

Having obtained mESCs expressing the *Ago2^{Halo}* allele from the endogenous Ago2 locus, we next set to test whether this genetically engineered allele could be used to map miRNA-mRNA interactions in cells. For these experiments, we adapted the HITS-CLIP method originally developed by the Darnell group (Chi et al., 2009) with two significant streamlining modifications enabled by the covalent bond between Halo-Ago2 and the HaloTag ligand. First, instead of using anti-Ago2 antibodies to isolate Ago2-containing complexes, we used sepharose beads covalently linked to the HaloTag ligands. Second, the radiolabeling and SDS-PAGE purification step required in conventional CLIP protocols to purify RNAs bound to Ago2 were omitted and replaced by extensive washes followed by direct RNA extraction from beads. We refer to this method as Halo-enhanced Ago2 pulldown (HEAP) (**Figure 2.3**).

To enable library preparation, we added two RNA adaptors to the RNAs. The first ligation was between a 3' adaptor and RNA in the complex after a series of on bead purifications. A 5' adaptor was added after the RNAs were released and purified from beads. These RNAs were then reverse transcribed into cDNAs. Amplification of the cDNA library by PCR produced a 65-bp band, corresponding to Halo-Ago2-bound adaptor-ligated mature miRNAs, and a smear (75~200 bp),

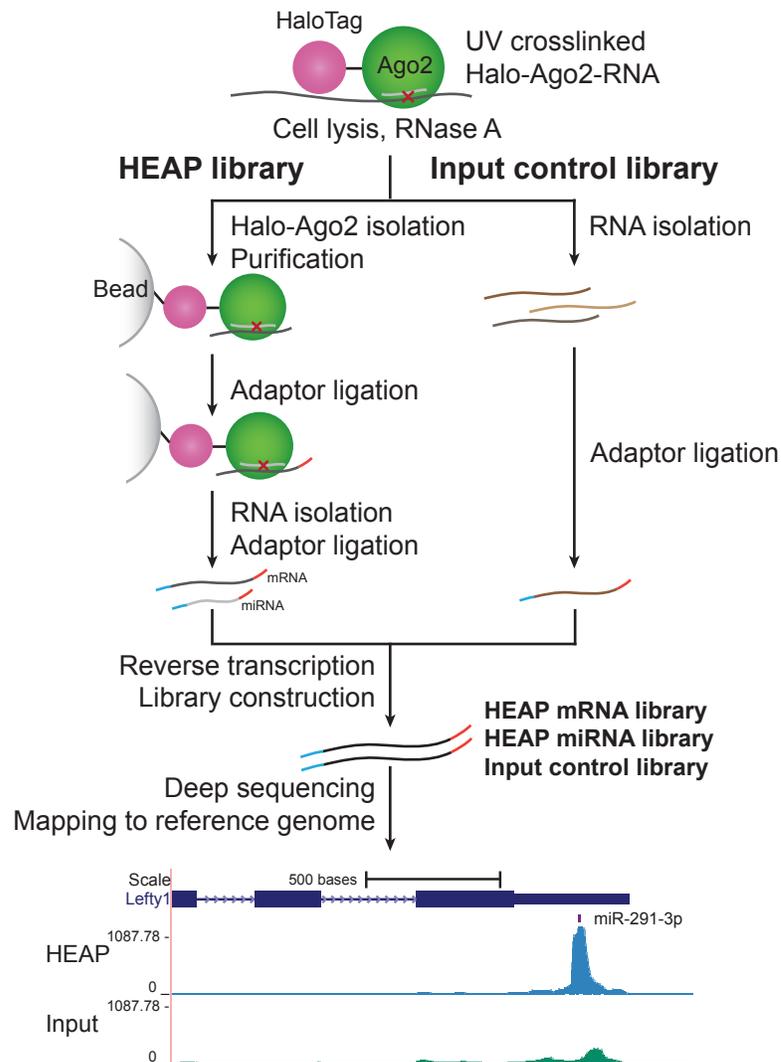


Figure 2.3: Schematic diagram of the HEAP method

Outline of the strategy used to generate HEAP and input control libraries (upper panel) and a representative Halo-Ago2 binding site identified in mESCs (lower panel).

predicted to contain the targets of these miRNAs (**Figure 2.4a**). The PCR products were resolved by PAGE and miRNAs and targets were purified separately. Afterwards, these libraries were subjected to high-throughput sequencing. While the miRNA target (mRNA) library allows the identification of miRNA binding sites, the miRNA library provides an estimate of miRNA abundance in the matching context

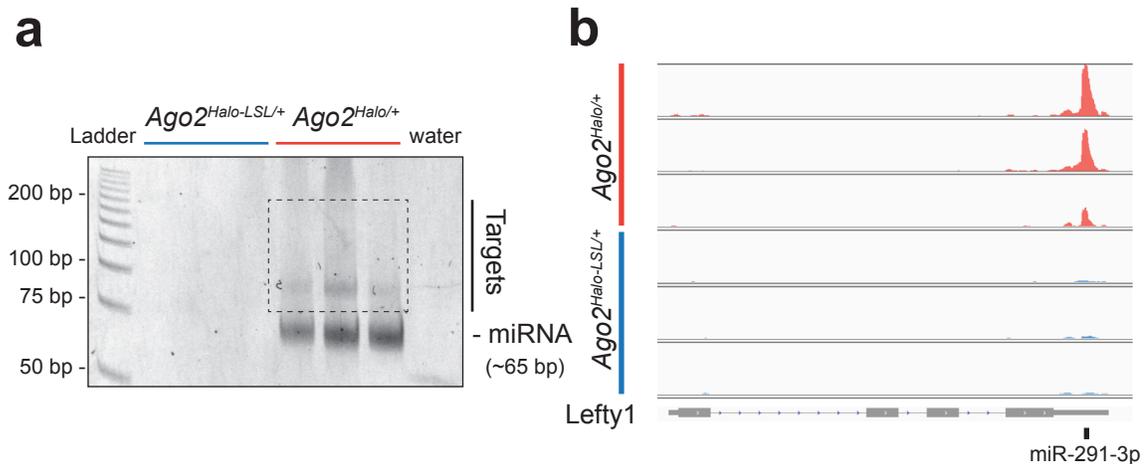


Figure 2.4: HEAP libraries generated from mESCs

a) Polyacrylamide gel electrophoresis of PCR products of HEAP libraries generated from *Ago2^{Halo-LSL/+}* and *Ago2^{Halo/+}* mESCs. A band corresponding to miRNA (~65 bp) and a smear corresponding to miRNA targets (75~200 bp, boxed) are visible in the *Ago2^{Halo/+}* samples. **b)** Genome browser view of *Lefty1* 3'UTR with tracks corresponding to HEAP mRNA libraries generated from *Ago2^{Halo/+}* (N = 3) and *Ago2^{Halo-LSL/+}* (N = 3) mESCs.

(Figure 2.3). The HEAP workflow was partially developed by Carla P. Concepcion.

In a first set of experiments, we generated three independent HEAP libraries from *Ago2^{Halo/+}* and *Ago2^{Halo-LSL/+}* mESCs. When mapped to the mouse standard genome (mm10), libraries generated from *Ago2^{Halo/+}* mESCs—but not those generated from control *Ago2^{Halo-LSL/+}* cells—produced well-defined “clusters” of reads, from here on referred to as peaks (Figure 2.4b).

To facilitate computational identification of these peaks, we generated control libraries by adapting the “SMinput” method used in eCLIP (Van Nostrand et al., 2016). The use of SMinput controls significantly improved specificity in the discovery of authentic RBP binding sites (Van Nostrand et al., 2016). We prepared input control libraries in parallel to the HEAP libraries by omitting the pulldown step. Size-matched PCR products (75~200 bp) were isolated by PAGE and processed identically to the HEAP libraries (Figure 2.3). Matching input control libraries were generated for every HEAP experiment discussed in this thesis.

Characterization of miRNA binding sites in mESCs

To map miRNA binding sites in mESCs, we generated three HEAP libraries and two input control libraries from *Ago2*^{Halo/+} mESCs (using 1.5×10^8 cells per library). Libraries were sequenced, and raw reads were processed. Only uniquely aligned reads to the standard mouse genome were used for peak discovery. We applied the CLIPanalyze package developed by our collaborator Yuri Pritykin (<https://bitbucket.org/leslielab/clipanalyze>) to identify peaks in these libraries. CLIPanalyze is an improved peak-calling algorithm based on edge detection technique similar to methods from image processing (Hsin et al., 2018; Lianoglou et al., 2013; Loeb et al., 2012). The algorithm merges all libraries (including HEAP mRNA libraries and input control libraries), smooths peak signals and defines a large set of putative peaks. Then it performs library size normalization using reads aligned to all exons outside of putative peaks. Finally, CLIPanalyze compares peak signal in the HEAP mRNA libraries and in the input control libraries and assigns FDR-corrected *p*-value to each peak (See **Method and APPENDIX II** for details). Preprocessing of the high-throughput sequencing data and peak calling were performed with the help of Yuri Pritykin and Yuheng Lu.

At an adjusted *p*-value cutoff of 0.05, CLIPanalyze identified a total of 30,564 Ago2 binding sites in our libraries. Previous studies have demonstrated that 3'UTRs of mRNAs are the preferred, although not exclusive, sites of interaction between miRNAs and mRNAs (Bartel, 2018). Consistent with these findings, the majority of HEAP peaks we identified in mESCs mapped to 3'UTRs, followed by sites mapping to CDS (**Figure 2.5**). The fraction of 3'UTR and CDS peaks increased monotonically with their statistical significance, while intergenic and intronic peaks had the opposite behavior. For example, when examining the 1,000

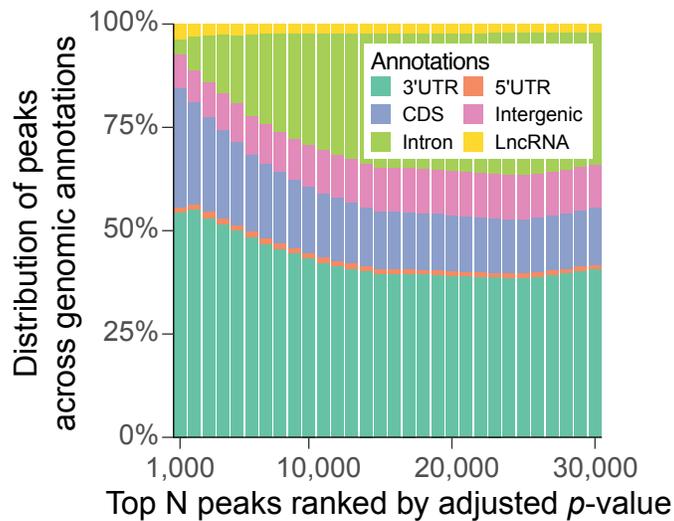


Figure 2.5: Genomic distribution of Halo-Ago2 binding sites

Peaks identified in the HEAP libraries from mESCs were ranked by increasing adjusted p -value before calculating their distribution across genomic annotations. CDS: coding sequence; 5'UTR: 5'-untranslated region; 3'UTR: 3'-untranslated region; LncRNA: long non-coding RNA.

most statistically significant peaks, greater than 50% of them mapped to 3'UTRs and less than 3% mapped to introns (**Figure 2.5**).

To gain additional insights into the nature of peaks identified by HEAP, we searched for enriched 7-nt motifs (7-mers) in the sequences underlying the 3'UTR peaks (**Figure 2.6a**). A series of overlapping 7-mers were identified when comparing frequencies of motifs in 3'UTR sequences within and outside of the peaks. To visualize these motifs, we mapped the enriched 7-mers back to genome and extracted sequences within a 15-nt window. These 15-nt motifs, along with background motifs, defined as 15-nt windows 100- or 200-bp away from the query motifs, were subjected to the HOMER *de novo* motif discovery algorithm (Heinz et al., 2010) for visualization. As a result, two 8-nt motifs were identified, which were complementary to the seed sequences of several miRNA families, including miR-291-3p, miR-292a-5p, miR-17-5p and miR-25-3p (**Figure 2.6b**). Moreover, we calculated log₂-fold enrichment of seed matches for all broadly conserved miR-

NA families in the 3'UTR peaks compared to background sequences. As shown in **Figure 2.6c**, seed matches for miR-291-3p, miR-17-5p and miR-25-3p were ranked top. miR-291-3p, miR-17-5p and miR-25-3p were among the most highly expressed miRNA families in mESCs, according to miRNA abundance estimated

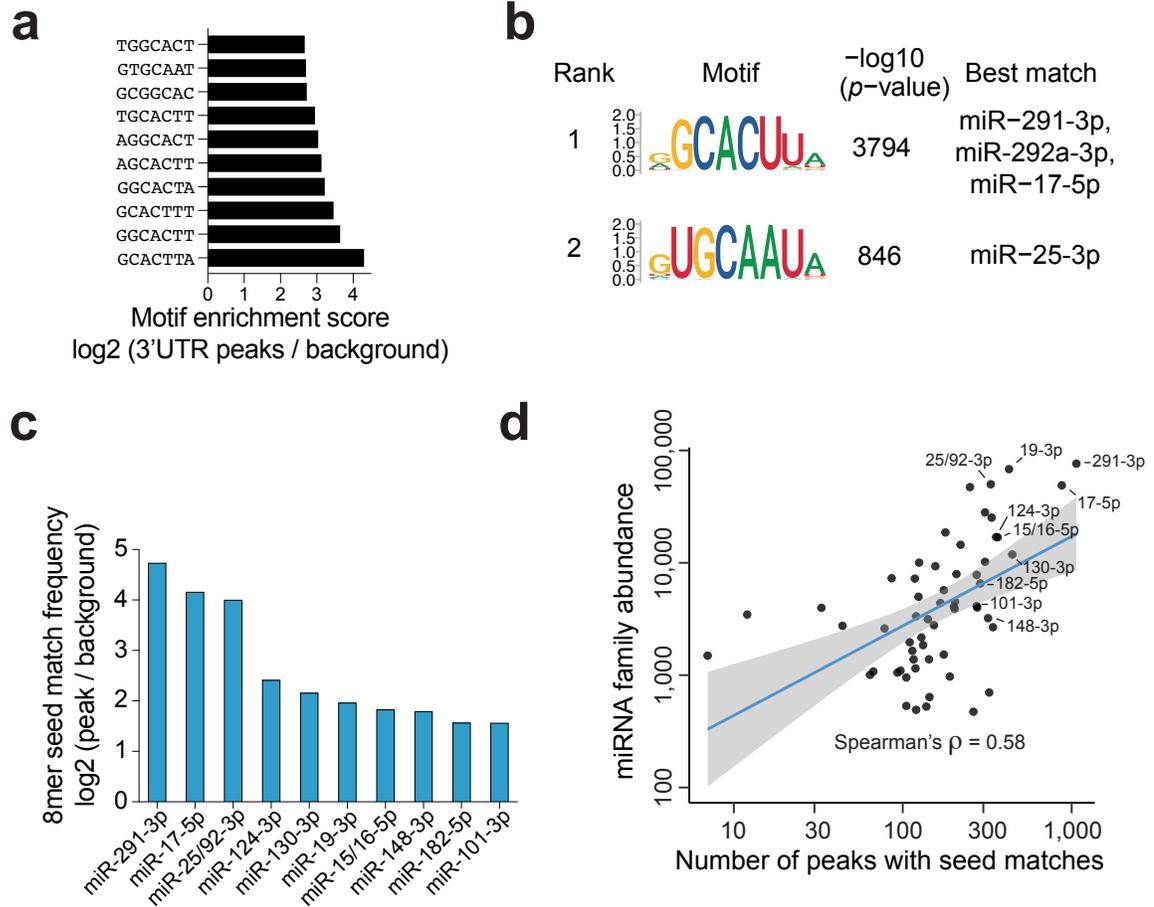


Figure 2.6: Enrichment of miRNA seed matches in HEAP 3'UTR binding sites
a) Bar plot showing the top ten most enriched 7-mers in Halo-Ago2 3'UTR peaks as determined by unbiased motif enrichment analysis. **b)** Sequence logos of the most enriched motifs in Halo-Ago2 3'UTR binding sites as determined by the HOMER *de novo* motif discovery algorithm. miRNA families whose seeds are complementary to these motifs are annotated. **c)** Enrichment for sequences complementary to murine miRNA seeds (8mers). The bar plot shows enrichment for the top ten miRNA seed families. **d)** Scatter plot showing the correlation between number of 3'UTR peaks with 7mer or 8mer seed matches to individual miRNA families and abundance of their corresponding miRNAs as measured in HEAP miRNA libraries. Blue line: best fit linear regression, with 95% confidence interval in grey.

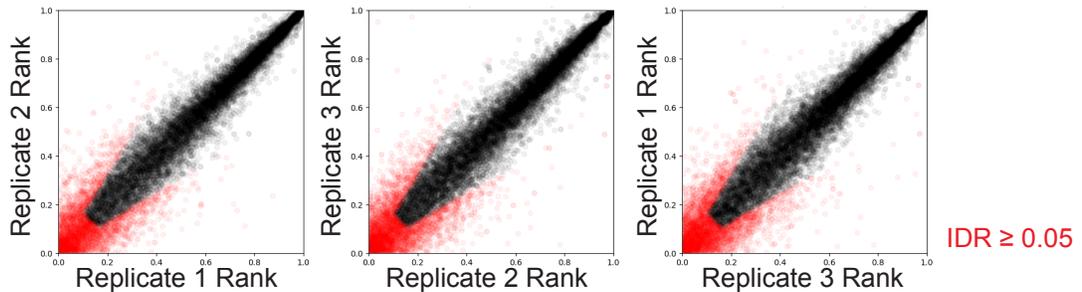
in the HEAP miRNA libraries.

To gain a better understanding of the relationship between miRNA expression levels and target abundance, we assigned the 3'UTR binding sites to a set of highly expressed and broadly conserved miRNA families by searching for 7mer (7mer-m8 or 7mer-A1) or 8mer seed matches for the corresponding miRNA family in the sequences under these peaks. We observed a positive correlation between the relative abundance of individual miRNA families and the number of corresponding peaks identified by HEAP (**Figure 2.6d**).

Collectively, these results suggested that the HEAP method was able to identify biologically relevant miRNA targets in mESCs. The HEAP libraries also delineated a positive relationship between miRNA abundance and target abundance *in vitro*.

Reproducibility, sensitivity and detection limit of the HEAP method

To measure the reproducibility of the HEAP method, we applied the CLI-



Number and fraction of peaks passing an IDR cutoff of 0.05:
Replicate 1 vs. Replicate 2: 11902/14818 (80.3%)
Replicate 1 vs. Replicate 3: 13314/16625 (80.1%)
Replicate 2 vs. Replicate 3: 12879/16096 (80.0%)

Figure 2.7: Reproducibility of the HEAP method

Pairwise IDR analysis of three independent HEAP libraries generated from mESCs. Each point represents the rank of an individual peak as determined in each pair of the replicates. Points in red correspond to peaks that failed to pass an IDR cutoff of 0.05. Table below summarizes the absolute number and percentage of peaks that passed the 0.05 IDR cutoff in each comparison.

conserved 8-mer that is complementary to the miR-291-3p seed (**Figure 2.8b**). We used CRISPR-Cas9 and homologous recombination in mESCs to introduce point mutations designed to disrupt this seed match (Flemr and Buhler, 2015) (**Figure 2.8b**). Two clones homozygous for the desired mutations were generated using two different guide RNAs. HEAP libraries generated from both clones showed complete loss of the Lefty2 peak (**Figure 2.8c**). A global differential peak signal analysis visualized as volcano plot revealed a selective loss of this binding site in the *Lefty2* mutant cells (**Figure 2.8d**). The result demonstrated the ability of the HEAP method to map *bona fide* miRNA-RNA interactions in cells.

To determine the detection limit of the HEAP method, we generated a series of HEAP libraries using decreasing numbers of mESCs (from 1.5×10^8 to 1×10^3). Peaks were called using CLIPanalyze by comparing each library to the two input control libraries generated before. As expected, the total number of confidently

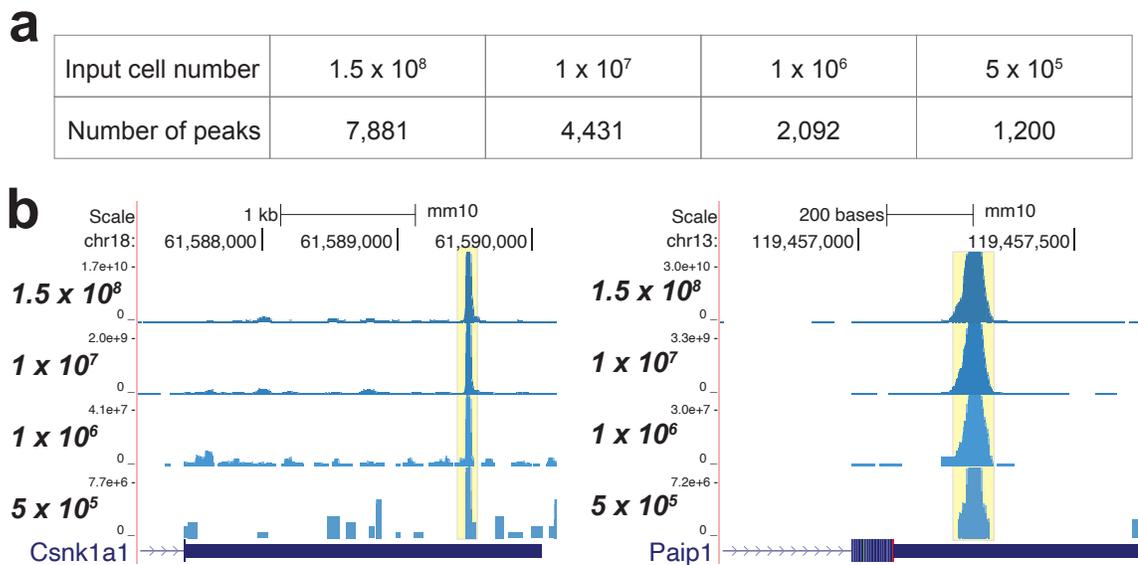


Figure 2.9: HEAP libraries generated from decreasing numbers of mESCs
a) Total number of peaks identified at an adjusted p -value cutoff of 0.1 in HEAP libraries generated from a series of numbers of mESCs. **b)** Genome browser view of representative peaks identified in HEAP libraries in panel a).

identified peaks progressively decreased as the amount of starting material was reduced (**Figure 2.9a**). Most robust peaks could be identified with as few as 5×10^5 mESCs, however, for optimal library resolution, a minimum of 1×10^7 cells was recommended (**Figure 2.9b**). Since mESCs have little cytoplasm, the detection limit is likely to be even lower for cell types with more abundant cytoplasm.

Collectively, these results indicate that the HEAP method is a robust and sensitive approach for miRNA binding site detection *in vitro*.

Gene repression associated with HEAP-identified miRNA binding sites

To assess the ability of HEAP to identify functional miRNA binding sites, we examined the expression changes of miRNA targets caused by miRNA pathway perturbation. We analyzed the RNA sequencing (RNA-seq) dataset generated by Bosson and colleagues from mESCs knockout for all four Argonaute proteins (*Ago1-4^{-/-}*) in the presence or absence of exogenously expressed Flag- and HA-tagged AGO2 (FHAGO2) [(Bosson et al., 2014), GSE61348]. To visualize the expression changes of miRNA targets, we plotted mRNA log₂ fold changes of selected gene sets between the FHAGO2 and control mESCs as cumulative distribution function (CDF) plot. CDF of all expressed genes (“All genes”) was used as dataset background. CDF curves of targets of selected miRNA families were plotted and their deviations from the background distribution were examined using two-sided Kolmogorov-Smirnov tests.

Introduction of FHAGO2 in *Ago1-4^{-/-}* cells should restore miRNA function, causing repression of their targets. In agreement with this prediction, miRNA targets containing HEAP 3'UTR binding sites for highly expressed miRNA families were preferentially repressed upon FHAGO2 reintroduction (**Figure 2.10a**). 3'UTRs with more than one miRNA binding site were under stronger suppression, indicative of

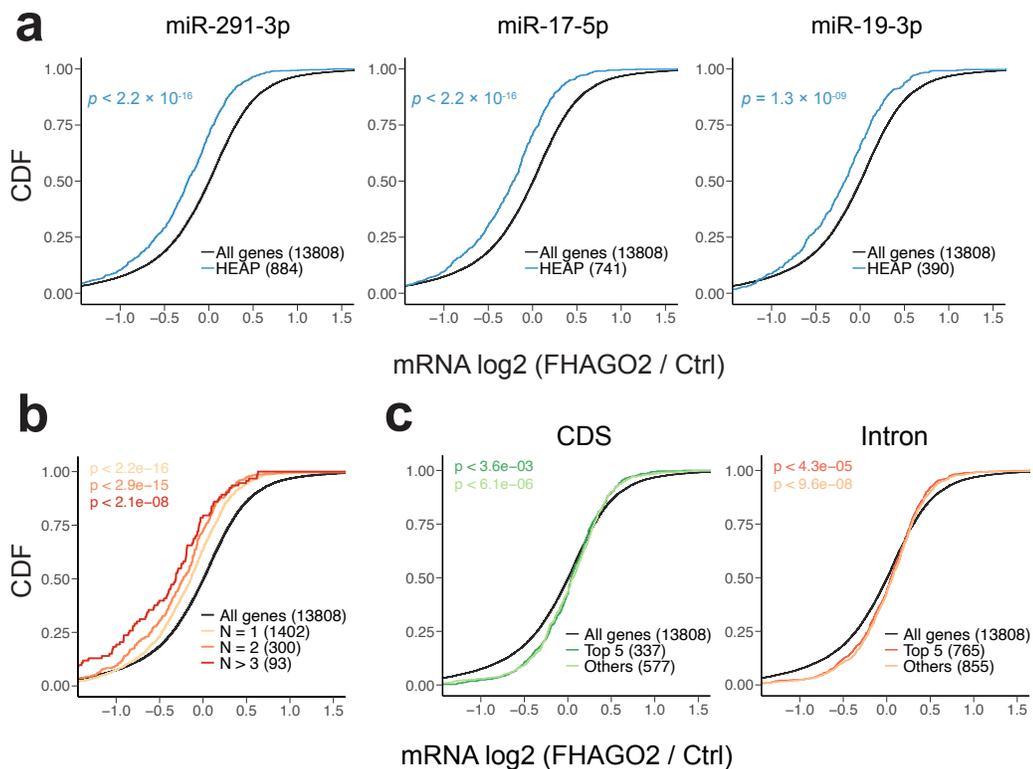


Figure 2.10: Gene repression associated with Halo-Ago2 binding sites

a) Cumulative distribution function (CDF) plots for HEAP-identified 3'UTR targets of selected miRNA families. The mRNA log₂ fold change was calculated in *Ago1-4^{-/-}* mESCs upon ectopic FHAGO2 expression. **b)** CDF plot for genes containing different numbers of HEAP peaks in their 3'UTRs. **c)** CDF plots for HEAP CDS and intronic targets. Peaks were grouped based on whether they contained seed matches for the highly expressed miRNA families (top 5). *P*-value: two-sided Kolmogorov–Smirnov test.

cooperation among miRNAs (**Figure 2.10b**). Conversely, and as expected, peaks mapping to other genomic annotations, such as CDS and introns, did not mediate consistent target repression (**Figure 2.10c**).

We quantified the amount of repression on targets of each miRNA family by calculating a z-score of the log₂ fold expression changes using the parametric analysis of gene set enrichment (PAGE) method (Kim and Volsky, 2005). A negative z-score suggests a preferential downregulation of targets. And the amplitude of the z-score is indicative of the strength of the repression. We found that targets

of most miRNA families had z-scores below zero. Furthermore, we observed a positive correlation between the absolute z-scores and relative abundance of miRNAs in corresponding families, consistent with the hypothesis that miRNA-target stoichiometry is a major determinant of gene repression (**Figure 2.11**).

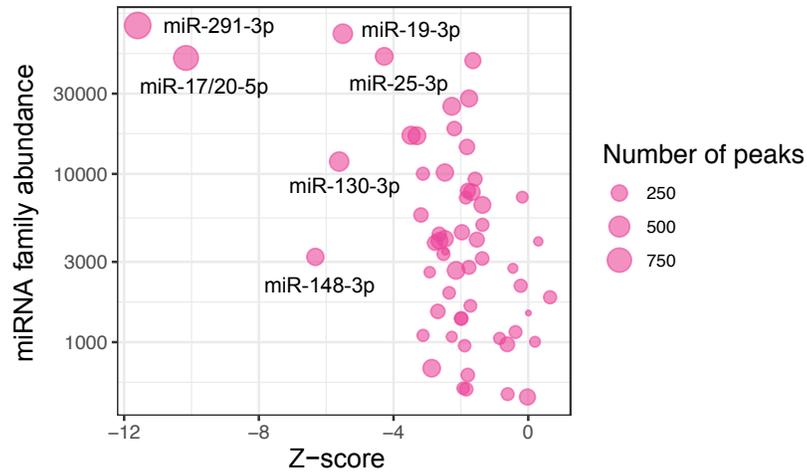


Figure 2.11: Correlation between miRNA abundance and gene repression capability

Scatter plot showing the correlation between target repression effects observed for the most abundant and broadly conserved miRNA families ($n = 58$) and miRNA abundance. Z-scores were calculated in targets containing 7mer or 8mer seed matches for the selected miRNA families using the parametric analysis of gene set enrichment (PAGE) method. miRNA abundance was calculated in HEAP miRNA libraries. The area of each circle is proportional to the number of targets identified for the corresponding miRNA family.

Next, we tested whether any characteristics of the peaks could serve as predictors for the repressive effects observed on their targets. In addition to adjusted p -values, log₂ fold change of peak signal between the HEAP and input control libraries was also calculated by CLIPanalyze during peak calling. We ranked the peaks by either their statistical significance (adjusted p -values) or log₂ fold changes and split them into quartiles. The first quartile corresponded to the most significant peaks or peaks with highest signal enrichments over input control libraries.

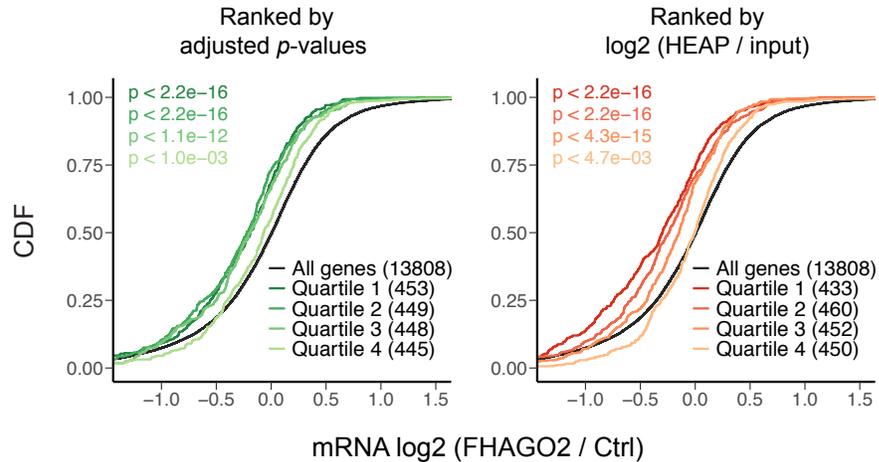


Figure 2.12: Benchmarking peak ranking parameters

CDF plots for HEAP targets of the 5 most abundant miRNA families. Left: peaks were ranked by adjusted p -values (HEAP vs. input control) and divided into quartiles; right: peaks were ranked by log2 fold change of peak signal (HEAP vs. input control) and divided into quartiles. P -values were calculated using two-sided Kolmogorov–Smirnov tests comparing each gene set to background (all genes).

We found that peaks with higher statistical significance (lower adjusted p -values) or higher log2 fold changes were associated with stronger target repression effects (Figure 2.12).

Collectively, by combining HEAP and CLIPAnalyze, we have identified a large class of functional miRNA binding sites. Furthermore, the adjusted p -value and log2 fold change (HEAP vs. input) generated by CLIPAnalyze provide a viable way to stratify these peaks based on their ability to repress targets.

Measuring translation inhibition associated with HEAP targets

miRNA-mediated gene repression manifests as mRNA degradation and/or translation inhibition. We have observed preferential mRNA repression in genes containing the 3'UTR peaks identified by HEAP. Here, we wanted to examine whether these miRNA binding sites could direct translation inhibition in addition to their ability to cause mRNA destabilization.

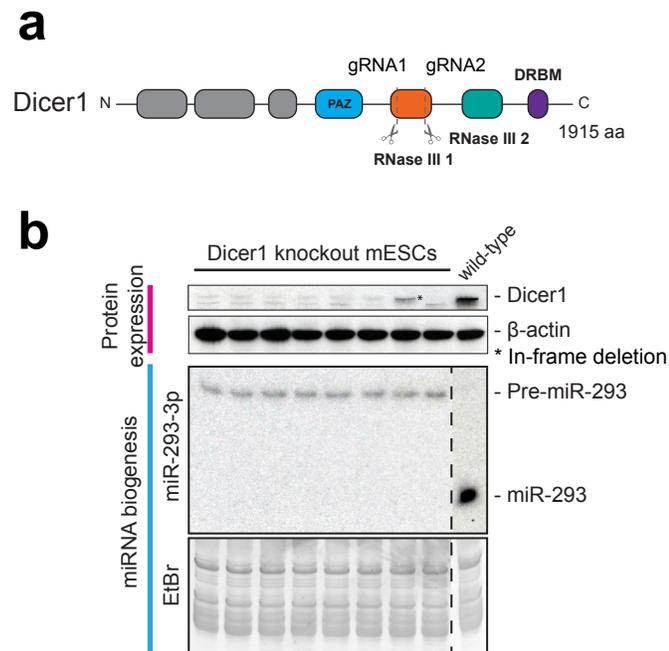


Figure 2.13: Generation and characterization of the Dicer1 knockout mESCs

a) Schematic structure of mouse Dicer1. A portion of the first RNase III domain was deleted using a pair of guide RNAs targeting the corresponding locations on the genome. PAZ: Piwi-Argonaute-Zwille. DRBM: double-stranded RNA binding motif. **b)** Dicer1 knockout single clones were characterized using immunoblot and northern blot. Upper panel: whole-cell lysates of mESC clones were probed with antibody against Dicer1 and β -actin. Lower panel: total RNAs were separated on 15% TBE-UREA PAGE, transferred to membrane and probed with oligonucleotides reverse complement to miR-293-3p. Vertical dashed line indicates splicing of the Northern blot membrane needed to align RNA and protein samples.

To enable simultaneous measurements of mRNA and protein expression changes caused by miRNAs, we generated mESCs lacking Dicer1. Deletion of Dicer1 should lead to blockade of maturation of most miRNAs and global de-repression of miRNA targets (Kanellopoulou et al., 2005). We knocked out *Dicer1* in *Ago2^{Halo/+}* mESCs using Cas9 and a pair of sgRNAs designed to delete a portion of the RNase III1 domain of Dicer1 (**Figure 2.13a**). Single clones harboring biallelic deletions were isolated. Western blot analysis confirmed the complete loss of Dicer1 protein in all clones. Northern blot for miRNAs revealed a depletion of

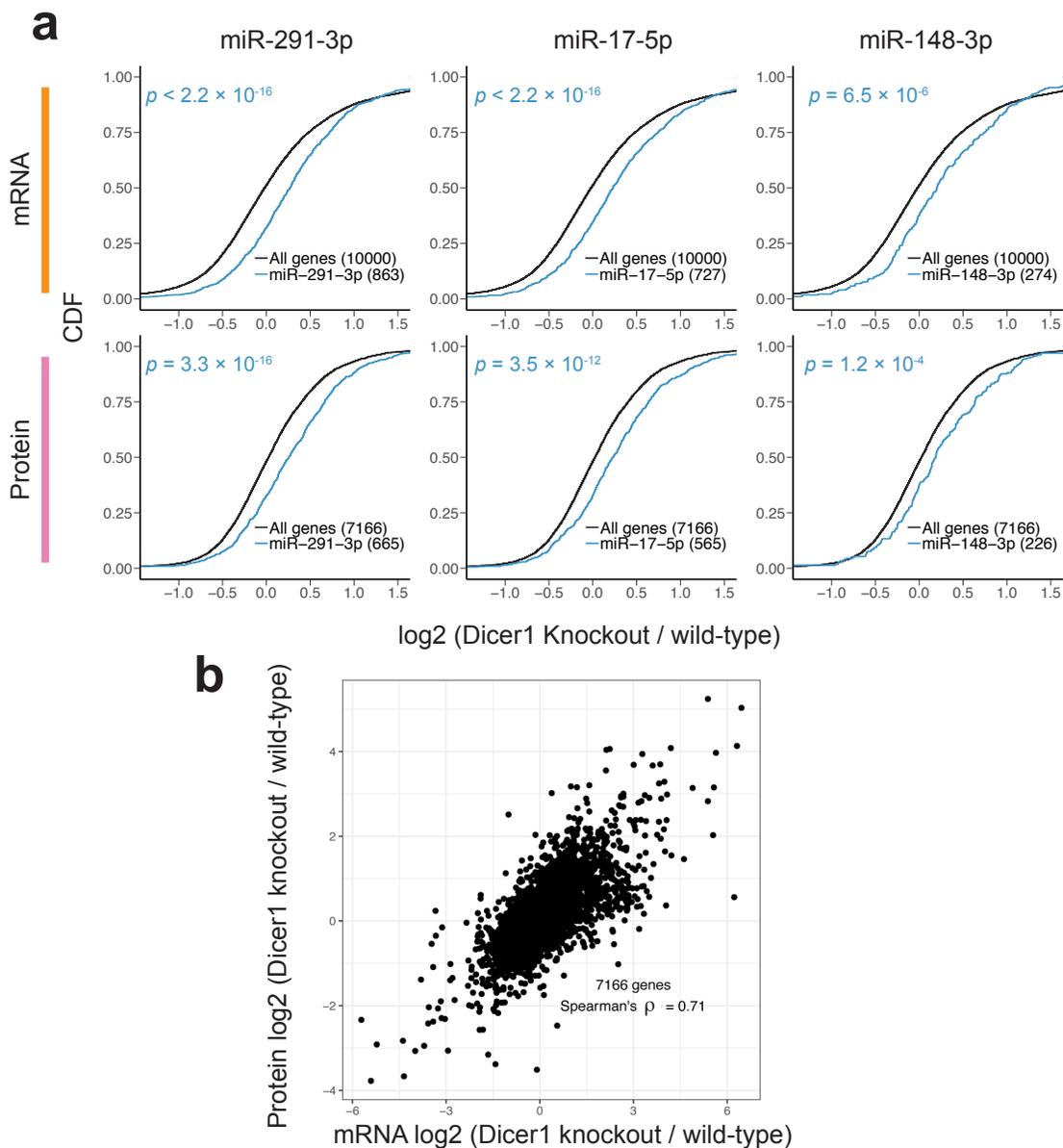


Figure 2.14: Correlation between protein and mRNA expression in Dicer1 knockout mESCs

a) CDF plots for targets of miR-291-3p, miR-17-5p and miR-148-3p identified by HEAP. The log₂-fold changes in mRNA and protein levels were calculated as Dicer1 knockout vs. wild-type mESCs. *P*-value: two-sided Kolmogorov–Smirnov test. b) Changes in mRNA expression between the Dicer1 knockout and wild-type mESCs highly correlate with changes in protein expression. mRNA expressions were measured by RNA-seq. Protein expressions were measured using Tandem Mass Tag Mass Spectrometry by Olesja Popow and Joao A. Paulo.

mature miR-293-3p, accompanied by an accumulation of pre-miR-293, suggesting a complete blockade of miR-293 maturation (**Figure 2.13b**).

The mutant mESC clones, along with wild-type clones, were submitted for transcriptomic and proteomic profiling by RNA-seq and Tandem Mass Tag Mass Spectrometry, respectively. Log₂-fold gene expression changes between the Dicer1 knockout and wild-type cells were calculated. 3'UTR targets identified in the HEAP libraries were preferentially upregulated at both mRNA and protein levels in the Dicer1 knockout mESCs (**Figure 2.14a**). Furthermore, there was an overall correlation between the mRNA and protein abundance (spearman's $\rho = 0.71$) (**Figure 2.14b**). Most of the protein expression changes observed could be explained by corresponding mRNA changes, in line with the hypothesis that mRNA destabilization is the dominant effect of mammalian miRNAs (Eichhorn et al., 2014; Guo et al., 2010). The proteomic analysis was performed by Olesja Popow and Joao A. Paulo at Harvard Medical School.

Comparison between HEAP and the conventional CLIP method

To benchmark the HEAP method, we compared miRNA target libraries generated by HEAP to those generated by one of the conventional Ago2 CLIP methods, iCLIP (Konig et al., 2010). We obtained the iCLIP dataset generated by Bosson and colleagues in *Ago1-4^{-/-}*-FHAGO2 mESCs (Bosson et al., 2014). In the iCLIP experiment, AGO2 binding sites were purified by tandem Flag-HA immunoprecipitation and a control library was generated using *Ago1-4^{-/-}* mESCs re-expressing wild-type AGO2. After read preprocessing, alignment and de-duplication, the iCLIP library was twice larger than each of the HEAP replicates, while the control libraries contained comparable number of reads (data not shown). Using peak-calling parameters optimized for the iCLIP method, CLIPanalyze algorithm

identified 6,813 FHAGO2 binding sites in the iCLIP library, and nearly twice as many (on average 13,532) in each of the three HEAP mESC libraries (nominal p -value < 0.1). The iCLIP library also identified significantly fewer peaks mapping to 3'UTR and more peaks mapping to intergenic regions compared to the HEAP libraries (**Figure 2.15a**).

3'UTR targets for the miR-291-3p seed family identified by both methods were associated with strong repression of the corresponding genes upon FHAGO2 reintroduction (**Figure 2.15c**). The overlap between miR-291-3p binding sites

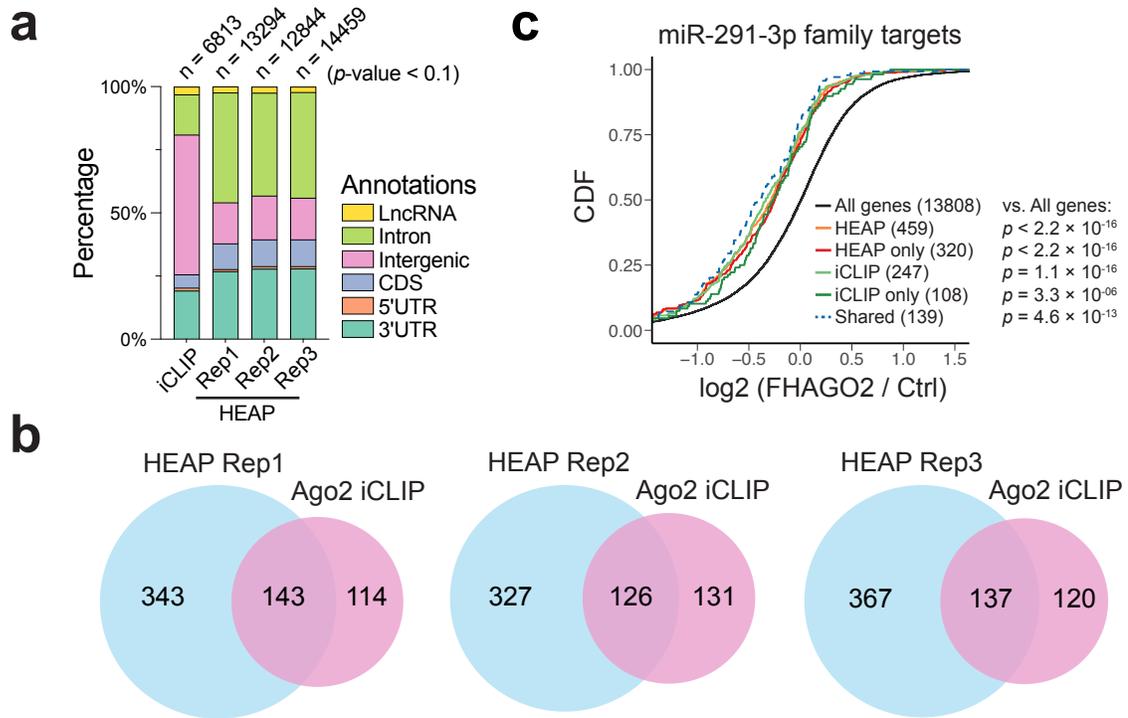


Figure 2.15: Comparison between HEAP and iCLIP

a) Genomic distribution of Ago2 binding sites identified in the iCLIP library and in the three HEAP libraries. The total number of peaks identified in each library is shown. **b)** Venn diagrams of the overlap between miR-291-3p 3'UTR binding sites identified by iCLIP and those identified by each HEAP library. **c)** CDF plot for targets of miR-291-3p identified by iCLIP or HEAP. The \log_2 fold change was calculated in *Ago1-4^{-/-}* mESCs upon ectopic FHAGO2 expression. HEAP only: targets identified uniquely by HEAP; iCLIP only: targets identified uniquely by iCLIP; shared: targets identified by both methods. P -value: two-sided Kolmogorov–Smirnov test.

identified by iCLIP and HEAP in 3'UTRs was partial, with the HEAP target pool being nearly twice as large (**Figure 2.15b**). Importantly, the targets identified only by HEAP also displayed strong repression upon FHAGO2 reintroduction, indicating that they are functional miRNA binding sites (**Figure 2.15c**).

Collectively, these results indicate that the HEAP method presents a sensitive and effective approach for functional miRNA target discovery *in vitro*.

Discussion

By applying the HEAP method, we have experimentally identified a large set of high-quality miRNA binding sites in mESCs. These binding sites were observed across genomic features. The fraction of peaks found in the introns declined as the peak-calling stringency increased, suggesting these binding sites were either less reproducible or had lower signal-to-noise. Peaks fell within the intergenic regions were ignored in our analysis.

The binding sites found in other genomic locations including CDS, 5'UTR and LncRNA resembled the 3'UTR peaks. Similar motifs were enriched in their sequences (data not shown), indicating miRNA-dependent Halo-Ago2 targeting. The functional consequences of miRNA binding in these regions remain largely unknown and controversial. For example, miRNA transfection experiments have shown that the CDS binding sites probably act by suppressing translation and have minimal effects on mRNA expression (Hausser et al., 2013; Zhang et al., 2018). In line with the latter observation is that the CDS sites identified by HEAP are not associated with mRNA repression (**Figure 2.10c**). However, two parallel experiments did not converge on the mechanism through which these CDS sites inhibited translation. One suggested that the CDS binding sites reduced ribosome occupancy on mRNAs (Hausser et al., 2013), while the other observed stalled ri-

bosomes in front of the CDS binding sites (Zhang et al., 2018).

We also benchmarked the HEAP method by comparing its performance to TargetScan7. In mESCs, the most abundant miRNA family is miR-291-3p. The HEAP method detected a larger number of binding sites for miR-291-3p than TargetScan (804 vs. 505) at the previously defined statistical cutoff. Binding sites unique to HEAP might be ignored by TargetScan due to low conservation of their sequences (**Figure 2.16a**). The overall ability of HEAP binding sites to repress targets was not as strong as that of TargetScan; however, it is important to point out that sites unique to HEAP also conferred statistically significant repression on their targets (**Figure 2.16b**).

Furthermore, we compared the HEAP libraries with a published Ago2 iCLIP

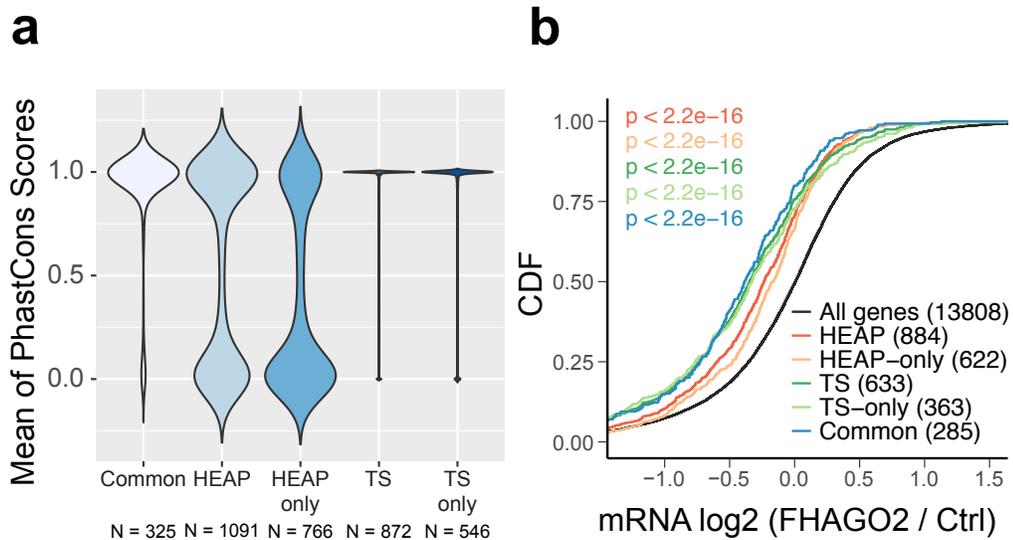


Figure 2.16: Comparison between miRNA binding sites identified by HEAP and TargetScan

a) Violin plot showing means of phastCons scores of miR-291-3p binding sites found by various methods. Common: miRNA binding sites found by both HEAP and TargetScan7; HEAP: sites identified by HEAP; TS: sites predicted by TargetScan7; HEAP-only: sites unique to HEAP; TS-only: sites unique to TargetScan7. N indicates the number of sites determined by each method. **b)** CDF plot for targets of miR-291-3p determined by the methods described above. *P*-values were calculated between each gene set to background (“all genes”) using two-sided Kolmogorov–Smirnov tests.

dataset. CLIPAnalyze identified 6,813 and on average 13,532 peaks in the iCLIP and HEAP dataset, respectively. It is worth noting that the parameters used for peak calling was optimized in the iCLIP dataset, which to certain degree compromised the performance of CLIPAnalyze in the HEAP libraries. While doing a favor for iCLIP, we still identified a larger set of peaks in the HEAP libraries at similar sequencing depth. Moreover, the HEAP peaks had higher signal-to-noise ratio than the iCLIP binding sites. Summits of HEAP peaks more consistently overlapped with seed matches for miRNAs. The improvement in peak resolution substantially enhances our ability to accurately assign putative miRNAs to these binding sites.

Last, the HEAP libraries revealed interesting miRNA-RNA interaction events in mESCs, which may be functionally important for mESC maintenance or even mouse embryonic development. One of such examples is the interaction between miR-291-3p and *Lefty2*, because *Lefty2* is involved in regulating mesoderm development as well as left-right patterning of mouse embryos (Meno et al., 1999; Meno et al., 2001). Using CRISPR-mediated gene editing and HEAP, we have validated this interaction *in vitro*. We are curious about the role of this interaction during development. Therefore, we generated a mouse strain carrying the same mutation (*Lefty2*^{MUT}) using CRISPR-mediated gene targeting. We are currently working on the phenotypic characterization of this strain. The generation of the *Lefty2*^{MUT} mice is described in **APPENDIX V**.

Method

Cell culture

Cells were maintained in a humidified incubator at 37 °C, 5% CO₂. mESCs were grown on irradiated MEFs in KnockOut DMEM (Gibco) supplied with 15% FBS (Gibco), leukemia inhibitory factor (Millipore, 10 U / mL), penicillin/streptomycin (Gibco, 50 U/ mL), GlutaMax (Gibco), non-essential amino acids (Sigma-Aldrich), nucleosides (Millipore) and 2-Mercaptoethanol (Bio-Rad, 100 μM). MEFs were cultured in DMEM (Gibco) containing 10% FBS, penicillin/streptomycin (100 U/mL) and L-glutamine.

Luciferase Assay

Ago2^{-/-} MEFs were transduced with the MSCV-PIG (Addgene: 21654), MSCV-PIG-Halo, MSCV-PIG-Halo-Ago2 or MSCV-PIG-Ago2 retroviruses to generate cell lines stably expressing HaloTag, the Halo-Ago2 fusion or Ago2. The dual-luciferase reporter assay system (Promega) was used to measure the cleavage activity of Halo-Ago2 and Ago2. Luciferase reporter plasmids pIS0 (*luc*⁺, Firefly luciferase, Addgene: 12178) (Yekta et al., 2004) and pIS1 (*Rluc*, Renilla luciferase, Addgene: 12179) were co-transfected into MEFs, along with a pSico vector expressing an shRNA against the Firefly luciferase or a control shRNA against CD8 (Ventura et al., 2008). The ratio between Firefly and Renilla luciferase activity was measured following manufacturer's instructions at 48 hrs after transfection.

mESC targeting

The targeting construct was generated by modifying the pKO-II vector through three steps of cloning. First, a fragment comprising a 2 kb 5' homology arm, the 5'UTR of Ago2, the HaloTag cDNA, the TEV protease recognition sequence, the coding sequence of Ago2 Exon1 and a portion of the first intron was

inserted into the pKO-II vector immediately upstream of the frt-PGK-NEO-frt cassette (between XhoI and AclI sites). Second, a 5 kb 3' homology arm was cloned into the HindIII site downstream of the frt-PGK-NEO-frt cassette. Lastly, a loxP-STOP-IRES-FLAG-loxP cassette was inserted into the AsiSI site between the TEV cleavage sequence and Ago2 coding sequence.

V6.5 mESCs (obtained from the Rudolf Jaenisch laboratory at Whitehead Institute and Massachusetts Institute of Technology) were electroporated with the linearized targeting construct and selected in mESC medium containing G418 for 7 days. Recombinant clones were identified by southern blot using probes designed against sequences outside the 5' and 3' homology regions. The *Ago2^{Halo-LSL/+}* and *Ago2^{Halo/+}* mESCs were derived from a validated clone by expressing and Flpe alone or in combination with Cre.

mESCs harboring the targeted allele were genotyped using a three-primer PCR (p1, 5'- GCAACGCCACCATGTACTC-3', final concentration 0.75 μ M; p2, 5'- GAGGACGGAGACCCGTTG-3', final concentration 1.0 μ M; p3, 5'-AGC-CGTTCCCTGAATCCTGTT-3', final concentration 0.5 μ M), which amplifies a 240-bp band from the wild-type allele (p1-p2), a 1281-bp band from the *Ago2^{Halo-LSL}* allele and a 651-bp band from the *Ago2^{Halo}* allele (p2-p3).

mESC mutagenesis

The Dicer1 knockout cells were generated from *Ago2^{Halo/+}* mESCs using CRISPR-Cas9. A pX333 vector (Addgene: 64073) (Maddalo et al., 2014) expressing Cas9 and a pair of guide RNAs designed to delete a portion of the RNase III 1 domain of Dicer1, was transiently transfected into *Ago2^{Halo/+}* mESCs. Single clones were isolated and genotyped by PCR.

The Lefty2 mutant clones were generated from *Ago2^{Halo/+}* mESCs using CRISPR-Cas9-mediated homologous recombination. PX330 vectors (Addgene:

42230) (Cong et al., 2013; Ran et al., 2013) expressing Cas9 and guide RNAs targeting the predicted miR-291-3p binding site in the 3'UTR of *Lefty2* were transiently transfected, together with single-stranded template DNAs, into *Ago2^{Halo/+}* mESCs. Clones undergoing homologous recombination were enriched using the method developed by Flemr and Buhler with plasmid pMB1610_pRR-Puro (Addgene: 65853) containing a fragment of guide RNA target sequence (Flemr and Buhler, 2015). Clones homozygous for the desired mutations were identified by PCR and Sanger sequencing.

Northern blotting

Total RNAs were isolated using TRIzol Reagent (Invitrogen) according to the manufacturer's instructions. 20 µg total RNAs from samples were loaded into a 15% TBE-Urea polyacrylamide gel and transferred to a Hybond-N⁺ membrane (GE Healthcare). After UV crosslinking and blocking, a ³²P-labeled DNA probe reverse complement to miR-293-3p was hybridized with the membrane at 37°C overnight. Next day, the membrane was washed and exposed to a film.

Mass spectrometry proteomics

Tandem Mass Tag mass spectrometry

Five independent Dicer1 knockout and five wild-type mESC clones were used in the proteomic analysis. Frozen cell pellets were lysed in 8 M urea and 200 mM EPPS, pH 8.5 with protease inhibitor (Roche) and lysates were additionally passed 10 times through a 21-gauge needle. Disulfide bonds were reduced using 5 mM tris(2-carboxyethyl)phosphine (30 min, RT) and alkylated with 10 mM iodoacetamide (30 min, RT in the dark). Alkylation reaction was quenched with 10 mM dithiothreitol for 15 min at RT. Per sample 100 µg protein (protein concentration determined prior to reduction/alkylation by BCA assay) were precipitated using methanol-chloroform precipitation and digested at RT with Lys-C protease (Wako

Chemicals) in 200 mM EPPS, pH 8.5 at a 50:1 protein:enzyme ratio overnight. More complete protein digestion was achieved through addition of trypsin (100:1 protein:enzyme ratio, Promega) for an additional 6 hrs at 37 °C. Acetonitrile was added to sample to a concentration of approximately 30%, and peptides were labelled with 0.2 mg TMT isobaric label reagent (Thermo Fisher Scientific) per sample for 1 hr at RT. Labelling reactions were quenched with the addition of hydroxylamine to 0.3% (v/v). Samples were combined at a 1:1:1:1:1:1:1:1:1 ratio and dried down by vacuum centrifugation. Excess TMT label was removed by C18 solid-phase extraction (Waters). The pooled sample was fractionated by off-line basic pH reversed-phase HPLC over a 50 min 5-35% acetonitrile gradient in 10 mM ammonium bicarbonate pH 8.0 into 96 fractions using an Agilent 300Extend C18 column (Wang et al., 2011). Collected fractions were combined into 48 fractions, of which 24 non-adjacent fractions were desalted using StageTips, dried by vacuum centrifugation and peptides were solubilized in 5% acetonitrile and 5% formic acid for subsequent LC-MS/MS analysis (Paulo et al., 2016). Approximately 2 µg of each sample was analyzed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) coupled to a Proxeon EASY-nLC 1200 liquid chromatography pump (Thermo Fisher Scientific) and a 100 µm × 35 cm microcapillary column packed with Accucore C18 resin (2.6 µm, 150 Å, Thermo Fisher). Peptides were fractionated over a 150 min gradient of 3 – 25% acetonitrile in 0.125% formic acid. An MS³-based TMT method was used, as described previously (McAlister et al., 2014; Paulo et al., 2016; Ting et al., 2011). MS¹ spectra were acquired with a resolution of 120,000, 350-1400 Th, an automatic gain control (AGC) target of 5e⁵, and a maximum injection time of 100 ms in the Orbitrap mass analyzer. The ten most intense ions were fragmented by collision-induced dissociation (CID) and analyzed in a quadrupole ion trap with AGC 2e⁴, normalized collision energy (NCE)

35, q-value 0.25, maximum injection time 120 ms, and an isolation window of 0.7 Th. MS³ spectra were acquired in the Orbitrap mass analyzer (AGC 2.5e⁵, NCE 65, maximum injection time 150 ms, 50,000 resolution at 400 Th) after fragmentation of MS² ions by HCD. Isolation windows were chosen depended on charge state z (z=2 1.3 Th, z=3 1 Th, z=4 0.8 Th, z=5 0.7 Th).

Mass spectrometry data processing

Spectra were searched using Sequest (Eng et al., 1994) with a 50 ppm precursor mass tolerance, 0.9 fragment ion tolerance and a maximum of two internal cleavage sites. Methionine oxidation was included as a variable modification, with a maximum of three modifications per peptide. Cysteine alkylation and TMT addition on lysines and peptide N-termini were set as fixed modifications. Spectra were searched against the Uniprot mouse proteome sequence database (downloaded on February 7th, 2014) containing both SwissProt and TrEMBL entries. Common contaminants were added to the database. The database was sorted in the following order: contaminant, SwissProt entries, TrEMBL entries and protein length within each category. All peptide sequences in the database were reversed and appended. FDR was estimated by linear discriminant analysis (Elias and Gygi, 2007; Peng et al., 2003), a 1% FDR filtering was applied at the peptide and protein level. Peptides were collapsed into a minimal number of protein identification as described by Huttlin and colleagues (Huttlin et al., 2010). This resulted in a filtered matrix of protein abundance values for 8,056 proteins. Then log₂FC of abundance was calculated for each protein by summing values within five replicates of each condition, adding 1 to each sum, and then taking log₂ of the ratio of the sums.

Halo-Ago2 live cell imaging

Ago2-null MEFs transduced with retroviruses MSCV-PIG, MSCV-PIG-Halo or MSCV-PIG-Halo-Ago2 were treated with 100 nM HaloTag TMRDirect ligand

(Promega) overnight and imaged on a ZEISS AXIO A1 microscope with AXIOCam MRC (ZEISS). A LD Plan-NEOFLUAR 20X/0.4 Ph2 korr objective was used.

Antibodies

Antibodies used were anti-Ago2 antibody (CST clone C34C6, 1:1000), anti- β -actin antibody (Sigma-Aldrich clone AC-74, 1:5000), anti-Tubulin antibody (Sigma-Aldrich clone DM1A, 1:5000), anti-HaloTag monoclonal antibody (Promega G9211, 1:1000), anti-Dicer1 (Bethy A301-936A, 1:1000) antibody and anti-TNRC6A (GW182) antibody (Bethyl A302-329, 1:1000).

RNA sequencing

Total RNAs from mESCs were extracted using TRIzol Reagent and subjected to DNase (Qiagen) treatment followed by RNeasy column clean-up (Qiagen). After quantification and quality control, 500ng of total RNA underwent poly(A) selection and TruSeq library preparation using the TruSeq Stranded mRNA LT Kit (Illumina) according to the manufacturer's instructions. Samples were barcoded and run on a HiSeq 2500 in a 50bp/50bp paired end run.

Reads were aligned to the standard mouse genome (mm10) using STAR v2.5.3a (Dobin et al., 2013). RNA reads aligned were counted at each gene locus. Expressed genes were subjected to differential gene expression analysis by DESeq2 v1.20.0 (Love et al., 2014).

Analysis of public datasets

The TT-FHAGO2 RNA-seq and iCLIP (GSE61348) (Bosson et al., 2014) datasets were processed and aligned to the UCSC mm10 mouse genome using STAR v2.5.3a. Reads mapping to multiple loci or with more than 5 mismatches were discarded.

miRNA targets z-score calculation

For conserved miRNA families, the mean log₂ fold change of predicted

targets compared to the rest of the transcriptome (background) was calculated. The means were converted to z-scores using an approach developed by Kim and Volsky (Kim and Volsky, 2005). $Z\text{-score} = (S_m - \mu) \times m^{1/2} / SD$, where S_m is the mean of log₂ fold changes of genes for a given gene set, m is the size of the gene set, and μ and SD are the mean and the standard deviation of background log₂ fold change values.

HEAP and input control library preparation

mESCs were harvested and irradiated with UV at dose 400 mJ/cm² in cold PBS on ice. Fresh tissues were harvested, homogenized and irradiated with UV for three times at dose 400 mJ/cm². Cell or tissue pellets were snap frozen on dry-ice and stored at -80 °C.

Frozen pellets were thawed, lysed with mammalian lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% Triton X-100 and 0.1% Na deoxycholate) containing protease inhibitor cocktail (Promega) and treated with RQ1 DNase (Promega) for 5 min at 37 °C. In order to get the “footprint” Halo-Ago2, lysates were treated with RNase A (Affymetrix, 1:50,000 diluted in TBS) for 5 min at 37 °C. ~2% of the lysates were saved for input control library preparation. The remaining lysates were diluted with buffer TBS (700 μL TBS per 300 μL lysates). For each sample, 300 μL Halolink resin (Promega) was used. The Halolink resin was equilibrated with TBS buffer containing 0.05% IGEPAL CA-630 and incubated with the TBS-diluted lysates at room temperature for 1.5 hr. After incubation, the resin was washed extensively with a series of buffers: SDS elution buffer (50 mM Tris-HCl, pH 7.5 and 0.1% SDS, one wash for 30 min at room temperature on a rotator), LiCl wash buffer (100 mM Tris-HCl, pH 8.0, 500 mM LiCl, 1% IGEPAL CA-630 and 1% Na deoxycholate, three times), 1× PXL buffer (1× PBS with 0.1% SDS, 0.5% Na deoxycholate and 0.5% IGEPAL CA-630, two times), 5× PXL buffer (5× PBS

with 0.1% SDS, 0.5% Na deoxycholate and 0.5% IGEPAL CA-630, two times) and PNK buffer (50 mM Tris-HCl, pH 7.4, 10 mM MgCl₂ and 0.5% IGEPAL CA-630, two times).

After dephosphorylation with calf intestinal alkaline phosphatase (Promega) at 37 °C for 20 min and washes with buffer PNK-EGTA (50 mM Tris-HCl, pH 7.4, 20 mM EGTA and 0.5% IGEPAL CA-630, two times) and PNK (two times), a 3' RNA adaptor with a phosphate on its 5' end (RL3) was ligated to the 3' end of RNAs using T4 RNA ligase 1 (NEB) at 16 °C overnight. Next day, the resin was sequentially washed with buffer 1× PXL (once), 5× PXL (once) and PNK (three times). RNAs on the resin were treated with T4 PNK (NEB) at 37 °C for 20 min and washed with buffer PNK (three times), Wash/Eq (once) and PK (100 mM Tris-HCl, pH 7.5, 50 mM NaCl and 10 mM EDTA, once). To release RNAs from the resin, proteins were digested with 4 mg/mL proteinase K (Roche) in PK buffer at 37 °C for 20 min and further inactivated by 7 M urea dissolved in PK buffer at 37 °C for 20 min. Free RNAs were extracted using phenol/chloroform and precipitated with ethanol/isopropanol at -20 °C overnight. Next day, RNAs were pelleted, washed with 70% cold ethanol and resuspended in DEPC-treated H₂O. A 5' RNA adaptor (RL5) with six degenerate nucleotides and a common 'G' on its 3' end (RL5-NN-NNNNG, RL5D-6N) was ligated to the purified RNAs using T4 RNA ligase 1 at 16 °C for 5 hrs. Then, the RNAs were treated with RQ1 DNase at 37 °C for 20 min to remove residual DNAs and purified by phenol/chloroform extraction and ethanol/isopropanol precipitation.

Purified RNAs were reverse transcribed using the DP3 primer (final concentration: 0.5 µM) and Superscript III reverse transcriptase (Invitrogen). The resulting cDNAs were amplified with primers DP3 and DP5 (final concentrations: 0.5 µM) and Accuprime Pfx DNA polymerase (Invitrogen) to the optimal amplification point. The

optimal amplification cycle (defined as the cycle before the PCR reaction reaching a plateau) was preliminarily determined by a diagnostic PCR visualized on gel or a real-time PCR with SYBR green (Invitrogen). PCR products of miRNAs (HEAP miRNA library, expected size: 65 bp) and targets (HEAP mRNA library, expected size range: 75~200 bp) were resolved on a 15% TBE-Urea polyacrylamide gel (Invitrogen) and extracted separately. To construct library for high-throughput sequencing, DNA primers DP3-barcodeX ("X" stands for barcode index) and DSFP5 (final concentrations: 0.33 μ M) containing Illumina adaptors, sequencing primer binding sites and Illumina TruSeq indexes for multiplexing were introduced to the HEAP miRNA and mRNA libraries by PCR. Sequencing libraries were run on a 6% TBE polyacrylamide gel (Invitrogen) and purified.

To prepare input control library, RNAs in the lysates saved before the HaloLink resin pulldown were dephosphorylated with calf intestinal alkaline phosphatase and phosphorylated using with T4 PNK. RNAs were then cleaned up using the MyONE Silane beads (ThermoFisher Scientific) as described in (Van Nostrand et al., 2016). Then, the 3' RNA adaptor (RL3) was ligated to the purified RNAs at 16 °C overnight. Next day, the ligated RNAs were purified using the MyONE Silane beads. Similar to the preparation of HEAP libraries, the RNAs were ligated to the 5' RNA adaptor (RL5D-6N) at 16 °C for 5 hrs, treated with RQ1 DNase, purified, reverse transcribed to cDNAs and amplified by PCR using primers DP3 and DP5. PCR products ranging from 75 to 200 bp were resolved on a 15% TBE-Urea polyacrylamide gel and used for input library preparation (see HEAP library preparation).

HEAP mRNA and miRNA libraries, along with the matched input control libraries, were submitted to the Integrated Genomics Operation Core at Memorial Sloan Kettering Cancer Center for high-throughput sequencing. After quantifica-

tion and quality control, libraries were pooled and run on a HiSeq 2500 in Rapid mode in a 100 bp or 125 bp single end run.

See also **Appendix I** for the step wise protocol of HEAP and **Appendix III** for oligos and adaptors used in library construction and sequencing.

HEAP library preprocessing

Barcode removal

The 6 nt degenerate barcodes and the last nucleotide 'G' coming from the 5' adaptor RL5D-6N (in total 7 nt) were removed from the beginning of reads and appended to the original read names, which later were used to distinguish duplicated reads produced at PCR amplification steps.

Adaptor removal and read quality control

The 3' adaptor (5'-GTGTCAGTCACTTCCAGCGGGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3') and bases with Phred quality score lower than 20 were trimmed from reads using cutadapt v1.15 or v1.17 (Martin, 2011). After trimming, reads shorter than 18 nt were discarded.

Alignment

Processed reads were aligned to the UCSC mm10 mouse genome using STAR v2.5.3a. Reads mapping to multiple loci or with more than 5 mismatches were discarded.

PCR duplicate removal

Reads mapped to the same locus with identical barcodes were considered PCR duplicates and therefore collapsed. This was achieved by storing aligned reads using chromosome names, strand information, positions of the first bases and the 7 nt barcodes as keywords. Representative reads of these unique events were written into a new BAM file, which was used for peak calling.

See also **Appendix II** for the code used in HEAP library preprocessing.

Peak calling

Peak calling was done using the unpublished package CLIPanalyze (<https://bitbucket.org/leslielab/clipanalyze>). The function `findPeaks()` was used to run multiple steps of analysis. First, the combined signal from uniquely aligned and PCR-duplicate-corrected reads from multiple replicates was convolved with the second derivative of a Gaussian filter. Zero-crossings of the convolved signal corresponded to edges of putative peaks. Second, read counting was run in putative peaks and in GENCODE-annotated gene exons with putative peaks subtracted, for both HEAP replicates and input control replicates. Library sizes for both HEAP and input control replicates were estimated using the read counts in exons outside of putative peaks. Third, using these library size estimates, differential read count analysis was performed between HEAP and input control read counts in putative peaks using DESeq2, and FDR-corrected p -values (adjusted p -values) were assigned to each peak. Peaks of size > 20 nt and read count $\log_2FC > 0$ in HEAP vs. input control were selected for downstream analysis. Peaks were annotated as overlapping with 3'UTR, 5'UTR, exons, introns, intergenic regions, lncRNA, in that order, using GENCODE (vM17) annotation. Peaks overlapping with genes of types "lincRNA", "antisense", "processed_transcript", according to GENCODE, were annotated as lncRNA peaks.

For mESCs, peak calling was run with the following parameters in `findPeaks()`: `count.threshold = 10`, `extend.slice = 10`, `bandwidth = 80`, `extend.peaks.in.genes = 150`. The full set of peaks was generated by comparing three independent HEAP libraries against two input control libraries. To identify peaks in each individual replicate to assess reproducibility and in the cell number titration experiment, a single library was compared to the two input control libraries. For iCLIP, peak calling was run using a single iCLIP library (TT-FHAGO2) against a single control

library (TT-AGO2) with the following parameters: count.threshold = 5, extend.slice = 50, bandwidth = 60, extend.peaks.in.genes = 150. For comparison with iCLIP, peak calling with the same parameters was run for each single HEAP library of comparable size against a single input control library.

See also **Appendix II** for the code used in peak calling.

miRNA abundance estimates

Reads in the HEAP miRNA libraries were processed and filtered following the “Barcode removal” and the “Adaptor removal and read quality control” steps described in the “HEAP library preprocessing” section. Processed small RNA reads were aligned to a miRNA genome index built from 1,915 murine pre-miRNA sequences from miRbase version 21 (Kozomara et al., 2018) (<ftp://mirbase.org/pub/mirbase/21/>) using Bowtie v2.3.4 (Langmead and Salzberg, 2012), and these reads were considered true miRNA counts if they fell within ± 4 bps at each of the 5' and 3' end of the annotated mature miRNAs. PCR duplicates were removed as described in the “PCR duplicate removal” step in the “HEAP library preprocessing” section.

miRNA seed family data were downloaded from the TargetScan website at http://www.targetscan.org/mmu_71/mmu_71_data_download/miR_Family_Info.txt.zip. For miRNA family level analysis, read counts mapping to members of the same miRNA family were summed up.

See also **Appendix II** for the code used in this analysis.

Motif discovery

Unbiased motif enrichment analysis

Frequencies of all k-nucleotide-long sequences (k-mers, $k = 7$) were calculated for sequences in selected peaks ($\text{Freq}_{\text{selected}}$) and background sequences (Freq_{bg}). The enrichment score for these 7-mers was calculated as $\log_2\text{FC} = \log_2$

$((\text{Freq}_{\text{selected}} + c) / (\text{Freq}_{\text{bg}} + c))$, where c was a small corrective value that depended on k , the number and size of peaks. K-mers with the highest $\log_2\text{FCs}$ were then reported. This analysis was performed using functions `calculateKmerBackground()` and `findKmerEnrich()` in `CLIPanalyze`. For mESCs, peaks mapping to 3'UTR were selected and background sequences were defined as sequences of 3'UTRs outside of peaks.

Enrichment score calculation for miRNA seed matches

\log_2 enrichment score of miRNA seed matches was calculated as $\log_2(\text{Freq}_{3'\text{UTR}} / \text{Freq}_{\text{bg}})$. $\text{Freq}_{3'\text{UTR}}$ was frequencies of 8mer seed matches for miRNA seed families in 3'UTR peaks, while Freq_{bg} was frequencies of these seed matches calculated in background sequences. Background sequences were defined as 3'UTR sequences outside of peaks.

HOMER de novo motif discovery

In mESC libraries, for the top 50 7-mers found by unbiased motif enrichment analysis, positions of their exact occurrences in 3'UTR peaks were found. Sequences of a 15-bp region around these occurrences were extracted and subjected to HOMER *de novo* motif discovery (Heinz et al., 2010), using 15-bp windows shifted 100 bp and 200 bp on both sides of the 7-mers (and excluding those overlapping with any of the 3'UTR peaks) as background sequences.

IDR analysis

IDR analysis was run using the python package at <https://github.com/nboley/idr> (Li et al., 2011). All putative peaks (size > 20 nt, $\log_2\text{FC} > 0$ for HEAP vs. input control) were provided via parameter "--peak-list". Peaks called for individual replicates were scored using $\log_2\text{FC}$ in HEAP vs. input control and provided via parameter "--samples", separately for each pair of replicates. Peaks at $\text{IDR} < 0.05$ were considered reproducible.

HEAP coverage analysis

bigWig files for visualization of HEAP and input control libraries at 1 bp resolution were produced in the following way. First, deepTools bamCoverage v3.1.3 (Ramirez et al., 2016) with parameter “-bs 1 --scaleFactor X” was used to produce bedGraph files. Here, size factors Y were estimated using DESeq2 applied to read counts in exons outside of peaks in all HEAP and input control libraries in a particular experimental model (mESCs, embryos, etc.) and then reciprocals $X = 1 / Y$ were used as BAM coverage scaling factors. Only bedGraph signal in the standard chromosomes was selected. Then “bedtools sort” (bedtools v2.23.0 (Quinlan and Hall, 2010)) and bedGraphToBigWig v4 (Kent et al., 2010) were used to produce bigWig files. HEAP libraries were visualized using UCSC genome browser or IGV (Robinson et al., 2011).

CHAPTER 3: *IN VIVO* IDENTIFICATION OF MIRNA-RNA INTER-ACTIONS USING THE HALO-AGO2 MOUSE STRAIN

Introduction

A growing body of evidence suggests that miRNA expression and regulation is cell type- and stage-dependent (Landgraf et al., 2007a). Comprehensive analysis of miRNAs in human and rodent tissues has revealed marked differences in miRNA expression profiles across tissues and cell types (Isakova et al., 2020; Lagos-Quintana et al., 2002; Liu et al., 2004a). Such differences are also observed across developmental and differentiation stages and between malignant and normal tissues (Chen et al., 2004; Krichevsky et al., 2003; Landgraf et al., 2007a; Lu et al., 2005). Therefore, the accurate identification of miRNA targets *in vivo* and in a cell-type-specific context is essential to dissect the roles of miRNAs in development, homeostasis, and disease.

miRNA target purification in a cell-type-specific manner has been demonstrated in the central nervous system of mice. Malmevik and colleagues expressed a GFP-Ago2 fusion protein driven by a synapsin promoter in the hippocampal neurons via adeno-associated viral (AAV) vector delivery. Then they purified neuron-specific miRNA targets through RNA immunoprecipitation (RIP) with anti-GFP antibodies (Malmevik et al., 2015). Despite the success in identifying miRNA-target interactions relevant to key neuronal phenotypes, this strategy has several limitations. First, ectopic expression of GFP-Ago2 may fail to recapitulate the regulation of endogenous Ago2 in hippocampal neurons. Second, the GFP-Ago2 fusion is only transiently expressed in tissues, making it unsuitable for longitudinal

studies. Third, the broad application of this approach is limited by the availability of cell-type-specific promoters. Last, not all tissues are accessible by the currently available somatic delivery methods.

The HEAP method described in the previous chapter is ideally suited to overcome these limitations. Not only can the HaloTag in principle increase purification efficiency, but the conditional nature of the knock-in allele we have generated can be combined with any of the many cell-type- or tissue-specific Cre-strains available to restrict expression of the Halo-Ago2 fusion to the desired cell types *in vivo*.

To translate the HEAP method to an *in vivo* setting, we used the mESCs harboring the conditional Halo-Ago2 allele to generate genetically engineered mice, in which expression of the Halo-Ago2 allele is induced in response to Cre-mediated recombination. In this chapter I will provide experimental evidence that this strain can be used to characterize Halo-Ago2-bound miRNA targets *in vivo*.

Generation of the Halo-Ago2 knock-in mice

To generate a mouse strain carrying the conditional Halo-Ago2 knock-in allele, a validated targeted mESC clone ($Ago2^{Halo-LSL; frt-PGK-NEO-frt/+}$) was injected into blastocyst. The targeting construct contained a frt-PGK-NEO-frt cassette to enable antibiotic selection of mESCs *in vitro*. Thus, to remove the Neomycin selection cassette, germline-transmitted animals were first crossed to the β -actin-Flpe strain (Rodriguez et al., 2000) (performed by Carla P. Concepcion). Then, we crossed these animals ($Ago2^{Halo-LSL}$) to CAG-Cre mice (Sakai and Miyazaki, 1997) to delete the LSL cassette and induce ubiquitous expression of the endogenous Halo-Ago2 allele ($Ago2^{Halo}$).

PCR and western blot analysis in MEFs and western blot in tissues derived

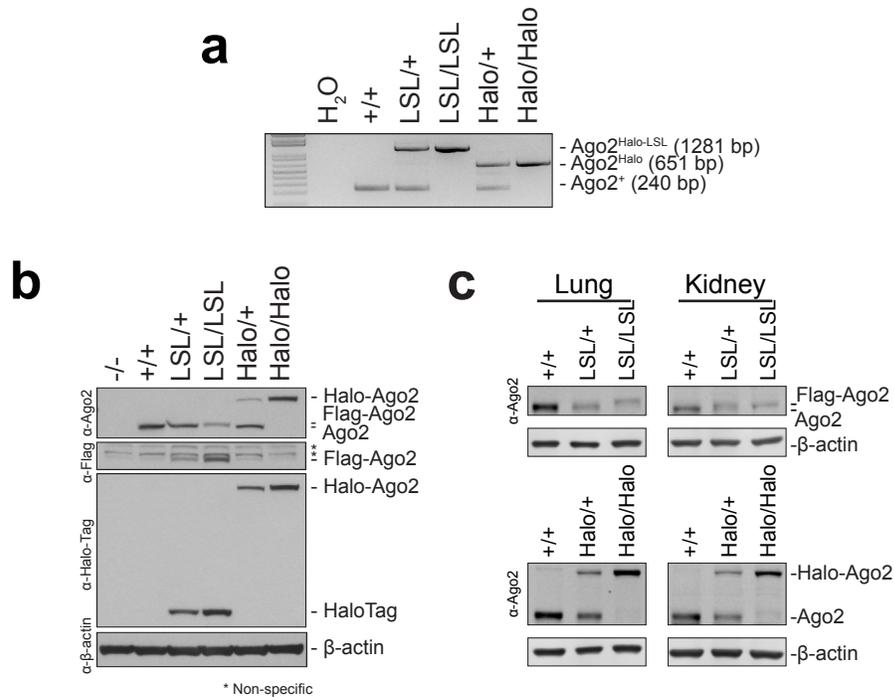


Figure 3.1: Genotyping and protein expressed from the Halo-Ago2 knock-in allele

a) Genotyping PCR in MEFs derived from E13.5 embryos harboring the indicated Ago2 alleles. +: wild-type Ago2; LSL: Ago2^{Halo-LSL}; Halo: Ago2^{Halo}. b) Whole-cell lysates of MEFs of indicated genotypes were probed with antibodies against Ago2, Flag, HaloTag and β-actin. -/-: Ago2^{-/-} MEFs. c) Expression of Ago2 fusion proteins in the lungs and kidneys of Ago2^{Halo-LSL} and Ago2^{Halo} mice.

from these mice confirmed efficient deletion of the LSL cassette and expression of the Halo-Ago2 protein (Figure 3.1).

Characterization of the Halo-Ago2 mice

Ago2 is essential for early embryonic development and postnatal survival (Liu et al., 2004b). To examine whether the presence of the HaloTag cassette had any adverse effect on mice, we performed heterozygous intercrosses in both Ago2^{Halo} and Ago2^{Halo-LSL} strains. Genotyping was done at weaning age. We found that although Ago2^{Halo/+} and Ago2^{Halo-LSL/+} mice were obtained at the expected Mendelian frequency and were phenotypically indistinguishable from wild-type mice,

homozygous mice for the *Ago2^{Halo}* or the *Ago2^{Halo-LSL}* alleles were recovered at sub-Mendelian frequencies (9.9% and 11.9%, respectively, compared to the expected 25%, **Table 3.1**). A reduced recovery of homozygous animals was also observed at birth, indicating that a subset of homozygous animals for either the recombined or unrecombined Halo-Ago2 allele were lost in utero.

Genotype	Wild-type (25% expected)	Heterozygous (50% expected)	Homozygous (25% expected)	<i>p</i> -value
<i>Ago2^{Halo-LSL}</i>	37/126 (29.4%)	74/126 (58.7%)	15/126 (11.9%)	0.00315**
<i>Ago2^{Halo}</i>	29/91 (31.9%)	53/91 (58.2%)	9/91 (9.9%)	0.00358**

Table 3.1: Absolute numbers and frequencies of genotypes obtained from heterozygous intercrosses of *Ago2^{Halo-LSL/+}* or *Ago2^{Halo/+}* mice.
P-value: Chi-Squared test.

Cheloufi and colleagues have found that the target slicing activity of Ago2 is essential for the perinatal survival in mice (Cheloufi et al., 2010). They showed that mice harboring a catalytically dead Ago2 (*Ago2^{ADH}*) display aberrant erythropoiesis and anemia, which possibly result from a blockade of miR-451 production in fetal livers. Fetal liver is the site of hematopoiesis before blood production switched to bone marrows. miR-451 knockout mice phenocopy the erythropoietic defects observed in the *Ago2^{ADH}* mice (Jee et al., 2018; Patrick et al., 2010; Rasmussen et al., 2010; Yu et al., 2010). Hence, we hypothesized that the sub-Mendelian recovery of *Ago2^{Halo}* homozygous mice might reflect aberrant miR-451 biogenesis and erythropoiesis during late gestation stage.

To test this hypothesis, we collected *Ago2^{+/+}*, *Ago2^{Halo/+}* and *Ago2^{Halo/Halo}* embryos at embryonic day 17.5 (E17.5) and analyzed miR-451 biogenesis in fetal livers. After processed by the microprocessor complex in the nucleus, pre-miR-451

is subjected to Ago2 cleavage that converts pre-miR-451 into a 30-nt long intermediate product. Then the 3'-end of the intermediate is resected while staying associated with Ago2, resulting in formation of mature miR-451 (Cheloufi et al., 2010; Cifuentes et al., 2010; Yang et al., 2010) (**Figure 3.2a**). Using small RNA northern blot, we found that pre-miR-451 was completely cleaved by Halo-Ago2 and pre-miR-451 was undetectable (**Figure 3.2b**). This was in contrast to what was observed in the *Ago2^{ADH}* embryos by Cheloufi and colleagues, where miR-451 maturation was completely blocked (Cheloufi et al., 2010). The result indicated that Halo-Ago2 largely retains endonucleolytic activity. Nevertheless, inspection of the mature miR-451 isoforms highlighted a reproducible shift towards longer isoforms in embryos carrying the *Ago2^{Halo}* allele, which might reflect the steric hindrance presented by the N-terminal HaloTag to the effective trimming of miR-451 on its 3'-end. Thus, these results suggested that despite its reduced cleavage activity, the Halo-Ago2 fusion was able to process all expressed miR-451 in erythrocytes

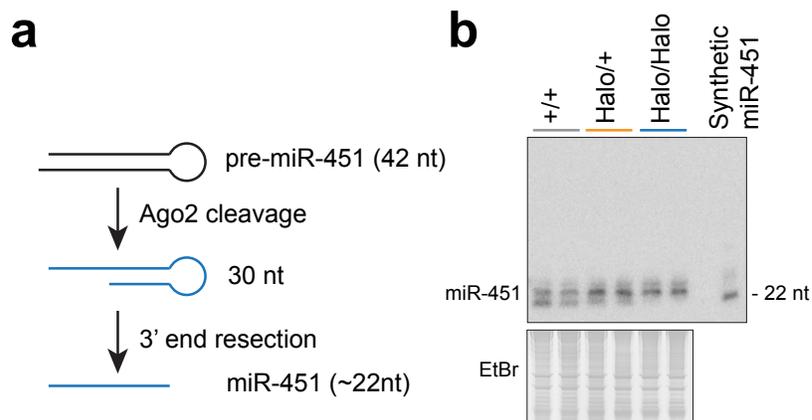


Figure 3.2: miR-451 isomiR shift in Halo-Ago2 expressing fetal livers

a) Schematic diagram of the Ago2-dependent miR-451 maturation pathway. **b)** Northern blot to examine the maturation and abundance of miR-451 in fetal liver cells. A synthetic miR-451 (22 nt) was loaded as a control. The membrane was probed with oligonucleotides reverse complement to mature miR-451. +/+ : *Ago2^{+/+}*; Halo/+ : *Ago2^{Halo/+}*; Halo/Halo : *Ago2^{Halo/Halo}*.

in fetal livers. Additional studies are needed to determine whether the shift in miR-451 isoforms is biologically consequential.

In parallel with the molecular characterizations, we also performed immunophenotyping of erythrocytes in the fetal livers of E17.5 embryos. Flow cytometry can be used to distinguish between four different stages of erythrocyte differentiation: ProE, EryA, EryB and EryC. Pro-erythroblasts (ProE) give rise to EryA erythroblasts (Ter-119^{high}, CD71^{high}, FSC^{high}), which then become smaller (EryB: Ter-

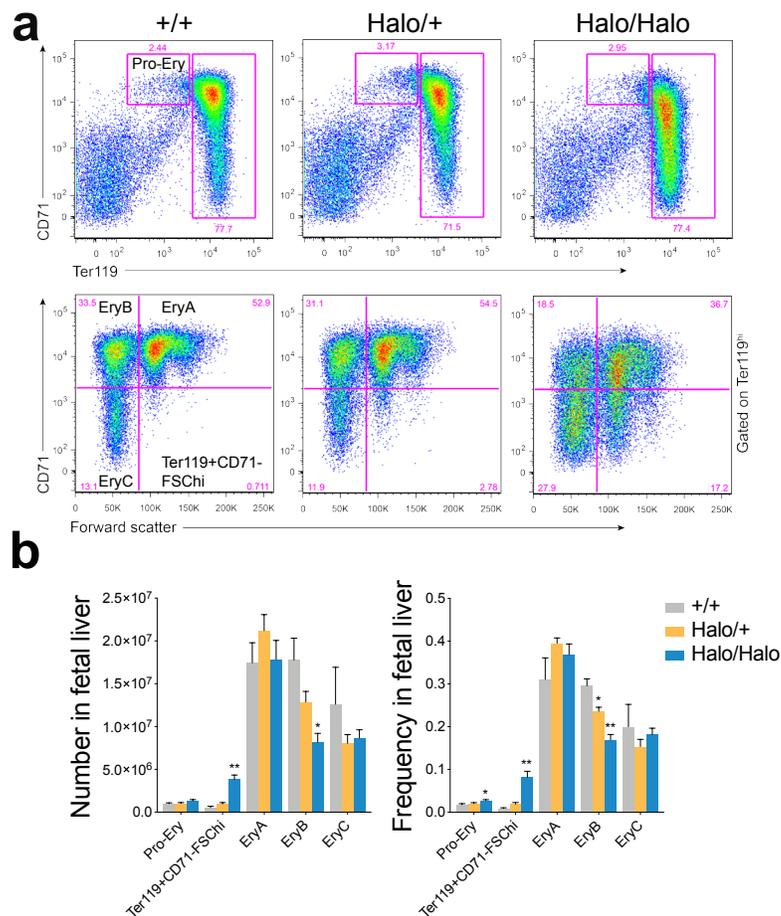


Figure 3.3: Aberrant erythropoiesis in Halo-Ago2 knock-in embryos
a) Flow cytometric analysis of fetal livers from *Ago2*^{+/+} (+/+), *Ago2*^{Halo/+} (Halo/+) and *Ago2*^{Halo/Halo} (Halo/Halo) E17.5 embryos. Cells were stained with anti-CD71 and Ter-119 erythroid lineage antibodies. **b)** Bar graphs summarizing absolute numbers and frequencies of erythroid sub-populations observed in **a)**. *P*-value: t-test. The analysis was performed by Bryan King.

119^{high}, CD71^{high}, FSC^{low}) and eventually reduce CD71 expression and differentiate into mature erythroblast (EryC: Ter-119^{high}, CD71^{low}, FSC^{low}) (Koulis et al., 2011). In *Ago2^{Halo/+}* and *Ago2^{Halo/Halo}* embryos, however, a new cell population defined as Ter-119^{high}, CD71^{low}, FSC^{high} accumulated, which was not detectable in wild-type embryos. This was accompanied by a significant dose-dependent loss of the EryB cell population (**Figure 3.3**). The new cell population might reflect erythroblasts that failed to undergo size reduction before losing CD71. Additionally, the erythroblast subsets were more heterogeneous in *Ago2^{Halo}* embryos as they were more dispersive along the Ter-119 and FSC axes. These evidence suggested aberrant erythropoiesis in the fetal liver of *Ago2^{Halo}* embryos and might explain at least in part the reduced frequency of viable homozygous mutant mice. The flow cytometry analysis of erythrocytes was performed by Bryan King, a postdoctoral fellow in the laboratory of Craig Thompson at Memorial Sloan Kettering Cancer Center.

Collectively, neither *Ago2^{Halo-LSL}* nor *Ago2^{Halo}* could completely replace wild-type *Ago2* in supporting mouse development. In the fetal liver of *Ago2^{Halo}* embryos, we observed abnormal erythroblast differentiation. Whether the abnormality is linked to the shift in miR-451 isoforms and whether the perinatal lethality can be attributed to the erythropoietic defects remain to be addressed.

Characterization of the Halo-Ago2 fusion protein

The sub-Mendelian recovery of homozygous mice might reflect lower *Ago2* expression levels compared to wild-type mice (**Figure 3.1b-c**) and/or an impaired miRISC formation or activity caused by the presence of the N-terminal tag. The phenotype observed in the *Ago2^{Halo}* mice highlighted the necessity of re-evaluating the activity of Halo-Ago2.

To determine whether Halo-Ago2 can be efficiently incorporated into

miRISC, we fractionated the crude lysates of *Ago2*^{Halo/+} MEFs using size-exclusion chromatography (La Rocca et al., 2015). Western blot analysis of the eluted fractions showed that the Halo-Ago2 fusion protein co-eluted with wild-type Ago2 in high molecular weight complexes, though a lower percentage of Halo-Ago2 was found in the large complexes compared to Ago2 (**Figure 3.4a**). Next, we examined whether Halo-Ago2 could interact with key components of the miRISC. We performed HaloTag pulldown from *Ago2*^{Halo/Halo} and *Ago2*^{Halo-LSL/Halo-LSL} MEFs using HaloTag ligands conjugated to magnetic beads. In the Halo-Ago2 allele, a short motif for TEV (tobacco etch virus) protease cleavage was placed between the

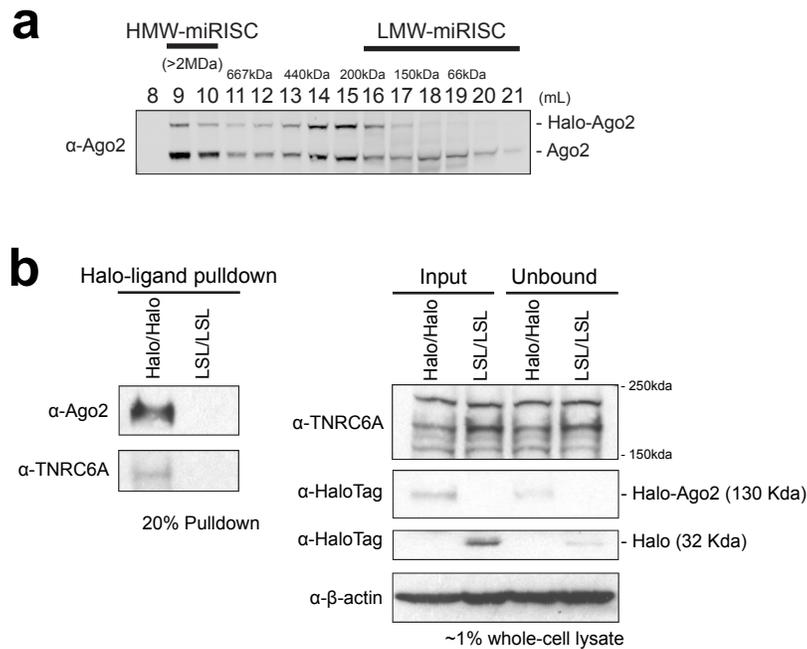


Figure 3.4: Incorporation of Halo-Ago2 into miRISC

a) Size exclusion chromatography fractionated whole-cell lysates from *Ago2*^{Halo/+} MEFs were probed with an anti-Ago2 antibody. HMW: high-molecular weight; LMW: low-molecular weight; miRISC: miRNA-induced silencing complex. The fractionation was performed by Minsi Zhang. **b)** HaloTag was pulled down from whole-cell lysates of *Ago2*^{Halo-LSL/Halo-LSL} (LSL/LSL) and *Ago2*^{Halo/Halo} (Halo/Halo) MEFs using HaloTag ligand conjugated to magnetic beads. Proteins co-purified were probed with antibodies against Ago2 and TNRC6A. Input and unbound lysates were probed with antibodies against TNRC6A, HaloTag and β -actin.

HaloTag and Ago2, which enables release of the complex after covalent pulldown. We treated the beads with TEV protease and analyzed the eluted proteins by western blot. GW182, the scaffold protein in the miRISC, was among the proteins purified from *Ago2^{Halo/Halo}* cells but not from *Ago2^{Halo-LSL/Halo-LSL}* cells, confirming the direct interaction between Halo-Ago2 and GW182 (**Figure 3.4b**). These results indicate that Halo-Ago2 can be assembled into miRISC and physically interacts with GW182; however, the interaction may be less stable compared to wild-type Ago2, contributing to explain the phenotype observed in the *Ago2^{Halo}* mice.

To more quantitatively define the effect of the presence of the HaloTag on miRISC activity and on RNAi, we used a series of reporter assays to measure miRNA-directed target repression in *Ago2^{Halo/Halo}* cells. First, we performed dual

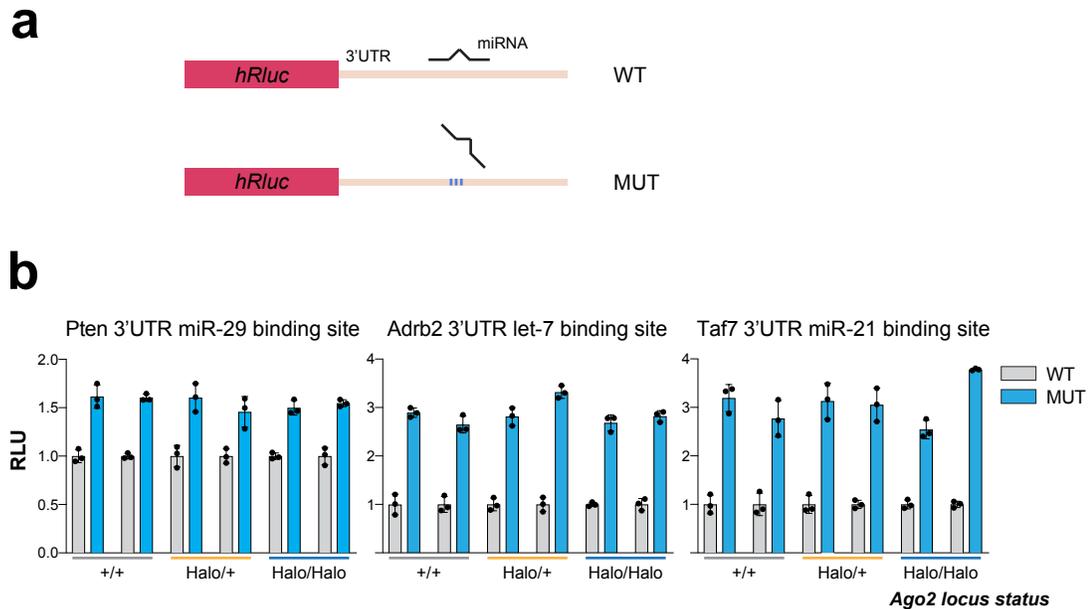


Figure 3.5: Efficient miRNA-mediated luciferase reporter repression in Halo-Ago2 expressing MEFs

a) Schematic of the luciferase reporter constructs used to measure miRNA-mediated gene silencing. **b)** Dual-luciferase reporter assays in *Ago2^{+/+}* (+/+), *Ago2^{Halo/+}* (Halo/+), and *Ago2^{Halo/Halo}* (Halo/Halo) MEFs. Experiment was performed in biological duplicates for each genotype. RLU: Relative luminometer units. Data are represented as Mean \pm SD.

luciferase reporter assays, in which two luciferase genes (Renilla and Firefly) were co-expressed from the same vector. 3'UTR fragments containing well-characterized miRNA binding sites of *Pten*, *Adrb2* and *Taf7* were cloned into the 3'UTR of the Renilla luciferase, while the Firefly luciferase served as an internal control. We also created control vectors by disrupting each of the miRNA binding sites (**Figure 3.5a**). We transiently transfected these reporter vectors into *Ago2*^{+/+}, *Ago2*^{Halo/+} and *Ago2*^{Halo/Halo} MEFs and measured the ratio between Renilla and Firefly luciferase activity 48 hours post-transfection. miRNA-mediated gene repression—measured as the ratio between mutant and wild-type reporters—was comparable across all genotypes (**Figure 3.5b**). This result suggested that, within the sensitivity of this assay, the amount of miRNA-mediated gene silencing was similar between *Ago2*^{+/+} and *Ago2*^{Halo/Halo} MEFs.

Next, we employed a more sensitive reporter assay, a two-color fluorescent reporter system (Mukherji et al., 2011), which measures miRNA-mediated target repression across a broad range of expression levels of the reporter and at the single cell level. The two-color fluorescent reporter contains a bi-directional tetracycline-responsive promoter that drives the expression of mCherry and eYFP. The mCherry expression cassette contains zero, one perfect match or four bulged binding sites for miR-20 in the 3'UTR (**Figure 3.6a**). The reporter with one perfect match site for miR-20 was included to measure the slicing activity of Ago2. Next, we transfected these reporters into *Ago2*^{+/+} and *Ago2*^{Halo/Halo} MEFs and measured fluorescent signals using flow cytometry 48 hours after doxycycline treatment. We binned the cells based on their eYFP signal and calculated the mean of mCherry intensity in each bin. As shown in **Figure 3.6b**, substantial repression was measured using both the bulged and perfect match reporters. Again, there was no detectable difference between Halo-Ago2 and Ago2 in directing miRNA-mediated

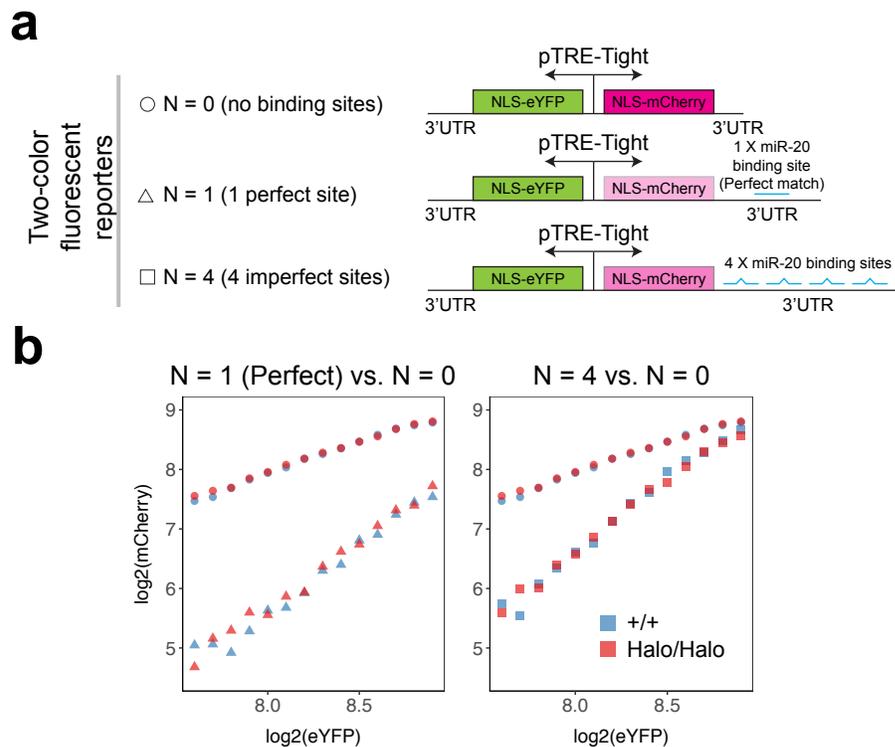


Figure 3.6: Efficient gene silencing and target slicing across a wide range of target expression levels in Halo-Ago2 expressing MEFs

a) Schematic of the two-color fluorescent reporters used. N indicates the number of miR-20 binding site(s) in the 3'UTR of mCherry. N = 0: the control reporter without miR-20 binding site; N = 1 (perfect): reporter containing 1 perfect binding site for miR-20; N = 4: reporter containing 4 imperfect binding sites for miR-20. **b)** Processed fluorescent signals measured by flow cytometry in *Ago2*^{+/+} (+/+) and *Ago2*^{Halo/Halo} (Halo/Halo) MEFs transfected with the indicated constructs.

repression and target cleavage, confirming that the Halo-Ago2 protein retains the ability to induce miRNA-mediated gene repression.

Finally, to systematically evaluate the transcriptomic changes caused by Halo-Ago2 expression, we performed RNA-seq in *Ago2*^{+/+} and *Ago2*^{Halo/Halo} MEFs. We reasoned that RNA-seq experiment would be more sensitive in measuring subtle expression changes that could have escaped the detection by both reporter assays. In this analysis, we included an RNA-seq dataset in MEFs expressing a T6B-YFP fusion as our positive control. The T6B protein corresponds to a portion

of the AGO interaction domain of human TNRC6B. Overexpression of T6B leads to dissociation of the miRISC and global de-repression of miRNA targets (by courtesy of Gaspare La Rocca, unpublished data). Gene expression changes were calculated as *Ago2*^{Halo/Halo} vs. *Ago2*^{+/+} or T6B vs. control. Z-scores for TargetScan predicted targets were calculated and plotted against the relative abundance of corresponding miRNA families. Expression of T6B induced a strong global de-repression of miRNA targets for almost every miRNA family. By contrast, only a modest upregulation was observed in *Ago2*^{Halo/Halo} MEFs for the most highly expressed miRNA families, such as miR-21-5p and miR-99-5p (**Figure 3.7**).

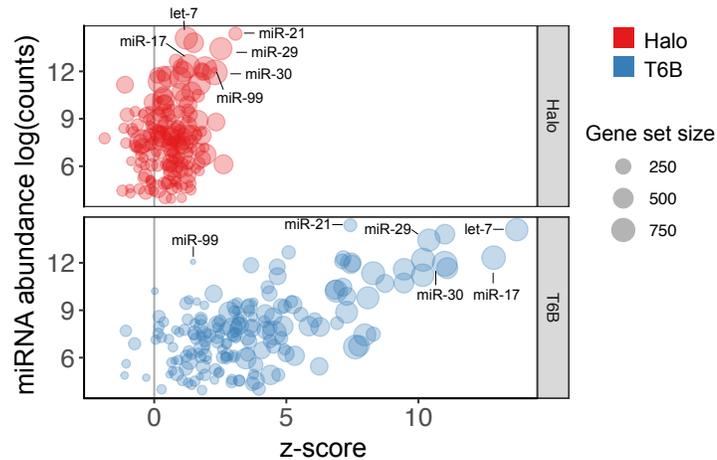


Figure 3.7: Gene expression profiling of Halo-Ago2 expressing MEFs

Scatter plots showing the effect of Halo-Ago2 expression on miRNA target repression and its correlation with miRNA abundance. TargetScan-predicted targets for all miRNA families were grouped into gene sets. Z-scores of mRNA log₂ fold changes of these gene sets in each dataset were calculated using the PAGE method and plotted on the x-axis. miRNA abundance was determined by small RNA sequencing and plotted on the y-axis. The area of each circle is proportional to the number of predicted targets for the corresponding miRNA family. Halo: *Ago2*^{Halo/Halo} vs. *Ago2*^{+/+}; T6B: *T6B*^{WT} vs. *T6B*^{MUT}.

In summary, the addition of HaloTag to Ago2 slightly impaired its ability to assemble into miRISC. Using reporter and gene expression profiling assays, small

transcriptomic changes were detected in *Ago2*^{Halo/Halo} MEFs, in which Halo-Ago2 was expressed at similar level to Ago2 in *Ago2*^{+/+} MEFs (**Figure 3.1b**). However, we cannot rule out the possibility that during embryogenesis the expression of Halo-Ago2 is lower or leads to severe phenotype in other cell types. It is also possible that the subtle perturbation of miRISC activity could be amplified during embryonic development and produce significant biological consequences.

Generation of HEAP libraries in vivo

The results presented in the previous section indicate that although the presence of the HaloTag may subtly impair Ago2 function, Halo-Ago2 is incorporated in the miRISC and is capable of promoting miRNA-mediated gene repression. Thus, the Halo-Ago2 knock-in mice we have developed can be utilized to map miRNA binding sites in live tissues. To benchmark the HEAP method *in vivo*, we generated libraries from the cortex of P13 mice, a tissue from which high-quality miRNA-target libraries had been previously generated by HITS-CLIP and CLEAR-CLIP (Chi et al., 2009; Moore et al., 2015). 2 HEAP libraries generated from the cortices of *Ago*^{Halo/+} mice produced 7,069 peaks at an adjusted *p*-value cutoff of 0.05. This number of miRNA-RNA interaction sites is comparable to that identified by Moore and colleagues (CLEAR-CLIP, GSE73059, *n* = 7,927) using 12 biological replicates (**Figure 3.8a-b**). HEAP and CLEAR-CLIP identified similar numbers of targets for miR-124-3p, one of the most abundant miRNAs in the mouse cortex (**Figure 3.8c**). When benchmarked against a microarray gene expression dataset generated from neuroblastoma cells (CAD) ectopically expressing miR-124 [(Makeyev et al., 2007), GSE8498], 2 HEAP libraries were as effective at identifying functional miR-124-3p sites as 12 CLEAR-CLIP libraries (**Figure 3.8d**). Collectively, these results demonstrate that the HEAP method provides a simple and

cost-effective approach to identify miRNA-RNA interactions in adult tissues.

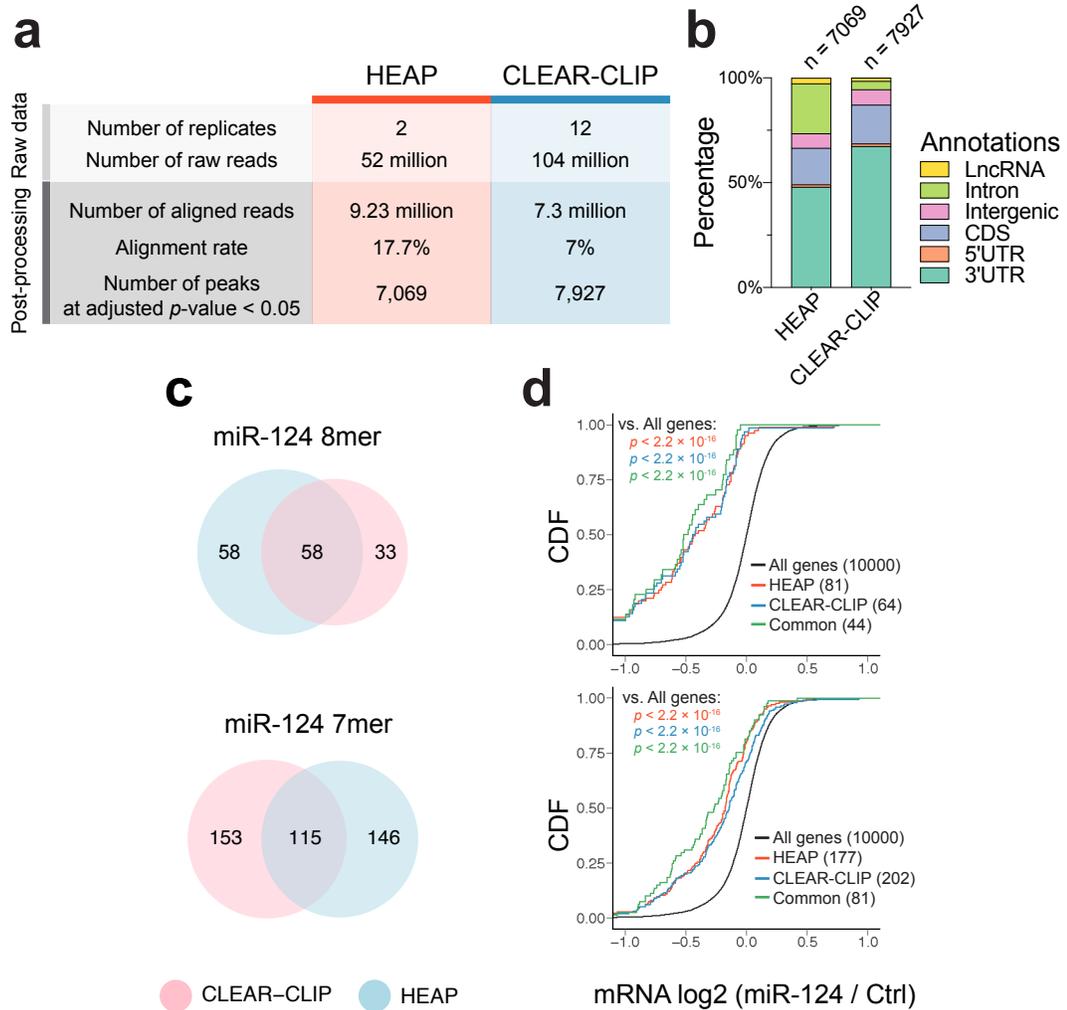


Figure 3.8: Comparison between HEAP and CLEAR-CLIP

a) Comparison between HEAP and CLEAR-CLIP libraries generated from cortices of P13 mice. b) Bar plot showing the genomic distribution of peaks identified by HEAP and CLEAR-CLIP. c) Venn diagrams showing the overlap between miR-124-3p target genes identified by CLEAR-CLIP and those identified by HEAP. Targets were grouped by seed matches types. d) CDF plots for targets of miR-124-3p identified by HEAP, by CLEAR-CLIP or by both methods (“Common”). Targets with 8mer (upper) or 7mer (lower) seed matches for miR-124-3p were plotted separately. mRNA log₂ fold changes were obtained from a dataset generated from a mouse neuroblastoma cell line (CAD) overexpressing miR-124. P -values were calculated between gene sets and background (“all genes”) using two-sided Kolmogorov–Smirnov tests.

Discussion

The fitness issue observed in both *Ago2^{Halo-LSL}* and *Ago2^{Halo}* mice emphasized the essentiality of the Ago2 protein in embryonic development and reflected the susceptibility of the locus to gene manipulation. The survival defects associated with the *Ago2^{Halo-LSL}* allele could be attributed to reduced Flag-Ago2 production from the engineered locus (**Figure 3.1b-c**). The insertion of the large loxP-STOP-IRES-Flag-loxP cassette to the first coding exon of Ago2 reduced its mRNA transcription (**Figure 3.9a**). Additionally, IRES seemed to be less efficient than the endogenous 5'UTR in driving translation (**Figure 3.1b-c**). By contrast, a paradoxical correlation between protein and mRNA produced from the *Ago2^{Halo}* allele was observed in MEFs and tissues (**Figure 3.1b-c**). RNA-seq revealed a dose-dependent decrease in Ago2 mRNA produced from the locus (**Figure 3.9b**). In *Ago2^{Halo/+}* cells, the Halo-Ago2 protein level was significantly lower than wild-type Ago2,

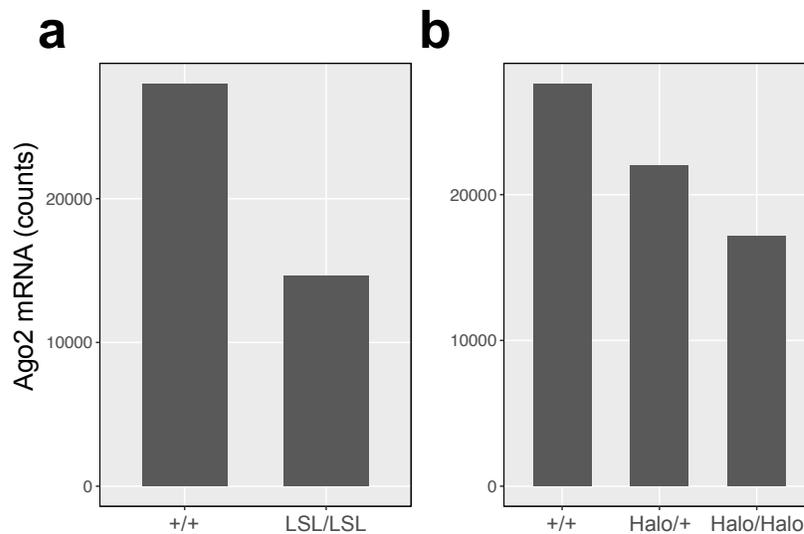


Figure 3.9: Reduced *Ago2* mRNA expression from the Halo-*Ago2* knock-in allele

a) Normalized counts of *Ago2* mRNA in *Ago2^{+/+}* (+/+) and *Ago2^{Halo-LSL/Halo-LSL}* (LSL/LSL) MEFs. **b**) Normalized counts of *Ago2* mRNA in *Ago2^{+/+}* (+/+), *Ago2^{Halo/+}* (Halo/+) and *Ago2^{Halo/Halo}* (Halo/Halo) MEFs.

whereas similar amount of Ago2 proteins were detected in *Ago2*^{+/+} and *Ago2*^{Halo/Halo} cells. The discrepancy might reflect preferential degradation of Halo-Ago2 in the presence of wild-type Ago2, or a translational feedback through which the cells upregulated Halo-Ago2 to compensate for its reduced activity.

Using biochemical approaches, we have demonstrated that Halo-Ago2 was slightly less effective compared to Ago2 in directing miRISC assembly. Inspection of the fetal livers suggested aberrant erythropoiesis. Nevertheless, it is worth noting that the perinatal lethality phenotype was ameliorated by maintaining the strains in a mixed genetic background (between C57BL/6J and 129S1). Given the fact that the *Ago2*^{Halo/+} and *Ago2*^{Halo-LSL/+} mice in the C57BL/6J background were phenotypically indistinguishable from their wild-type littermates, heterozygous animals can be easily generated and safely used for *in vivo* applications of the Halo-Ago2 knock-in mice.

Despite the caveats, the Halo-Ago2 knock-in allele offers several advantages over conventional strategies. First of all, the stable and physiologic expression of Halo-Ago2 permits miRNA targets discovery from virtually all tissue and cell types. The activation of Halo-Ago2 expression is controlled by the Cre recombinase, which can be introduced by either breeding or somatic gene delivery. By restricting Halo-Ago2 expression to cells of interest, miRNA targets can be purified even from hard-to-dissect tissues without exhaustive preliminary cell enrichment. Moreover, cells expressing Cre are irreversibly labeled with Halo-Ago2, allowing long-term management and continuous characterization of the cells. This is particularly useful for miRNA target identification in cells sharing a common progenitor, such as cells differentiated from the same stem cell population or tumor cells.

With the Halo-Ago2 knock-in mice, the biological roles of miRNAs can now be interrogated in a broader range of contexts *in vivo*. We envision the generation

of an atlas of miRNA binding sites across tissues and cell types under physiological and pathological conditions. The lineage-specific Halo-Ago2 activation and target purification can lead to better understanding of the roles of miRNAs in development and disease formation.

Method

Cell culture

All cells were maintained in a humidified incubator at 37 °C, 5% CO₂. MEFs were cultured in DMEM (Gibco) containing 10% FBS (VWR), penicillin/streptomycin (100 U/mL) and L-glutamine.

Mouse husbandry and generation of the conditional Halo-Ago2 knock-in mice

mESC targeting strategy was described in Chapter 1. A validated mESC clone harboring was injected into C57BL/6 blastocyst to generate chimeric mice. Mice heterozygous for the targeted allele were crossed to the β -actin-FIpe mice (Rodriguez et al., 2000) to remove the frt-PGK-NEO-frt cassette, resulting in the generation of *Ago2*^{Halo-LSL/+} mice. The *Ago2*^{Halo/+} mice were obtained by crossing the *Ago2*^{Halo-LSL/+} mice to the CAG-Cre mice (Rodriguez et al., 2000).

Mice carrying the knock-in alleles were genotyped as described before. All studies and procedures were approved by the Memorial Sloan Kettering Cancer Center Institutional Animal Care and Use Committee.

Northern blotting

Total RNAs were isolated using TRIzol Reagent (Invitrogen) according to the manufacturer's instructions. 20 μ g total RNAs from samples were loaded into a 15% TBE-Urea polyacrylamide gel and transferred to a Hybond-N⁺ membrane (GE Healthcare). A synthetic 22-nt miR-451 was loaded as control. After UV crosslinking and blocking, a ³²P-labeled DNA probe reverse complement to mature miR-451 was hybridized with the membrane at 37°C overnight. Next day, the membrane was washed and exposed to a film.

MEF derivation

MEFs were derived from mouse E13.5 embryos following standard protocols. *Ago2^{Halo}* and *Ago2^{Halo-LSL}* MEFs were generated by intercrossing *Ago2^{Halo/+}* and *Ago2^{Halo-LSL/+}* mice, respectively. MEFs were immortalized with retrovirus expressing the SV40 large T antigen (Zhao et al., 2003) (Addgene:13970).

Size exclusion chromatography

Cells were lysed with Sup6-150 buffer (150 mM NaCl, 10 mM Tris-HCl, pH 7.5, 2.5 mM MgCl₂, 0.01% Triton X-100, protease inhibitor (Roche) and phosphatase inhibitor (Roche)). Lysates were fractionated using the Superose 6 10/300 GL prepacked column (GE Healthcare) coupled with the AKTA FPLC system as described in (La Rocca et al., 2015; Olejniczak et al., 2013). Eluted proteins were concentrated by trichloroacetic acid (TCA)/acetone precipitation, analyzed by immunoblot and imaged using the Odyssey CLx imaging system (Li-Cor).

Isolation of Halo-Ago2/Tnrc6 complexes

Ago2^{Halo-LSL/Halo-LSL} and *Ago2^{Halo/Halo}* MEFs were lysed with HaloTag protein purification buffer (150 mM NaCl, 50 mM HEPES, pH 7.5, 0.005% IGEPAL CA-630) and lysates were incubated with HaloTag magnetic beads (Promega) for 90 min at room temperature on a rotator. After three washes with the HaloTag protein purification buffer, proteins were released by TEV protease (Invitrogen) digestion at 30 °C for 1 hr. Eluted proteins were analyzed by immunoblot and visualized using ECL (GE Healthcare).

Antibodies

Antibodies used were anti-Ago2 antibody (CST clone C34C6, 1:1000), anti-β-actin antibody (Sigma-Aldrich clone AC-74, 1:5000), anti-HaloTag monoclonal antibody (Promega G9211, 1:1000), anti-Flag antibody (Sigma-Aldrich F7425, 1:1000) and anti-TNRC6A (GW182) antibody (Bethyl A302-329, 1:1000).

Dual luciferase reporter assay

Fragments of the 3'UTRs of Pten and Adrb2 containing miRNA binding sites for miR-29-3p and let-7-5p, respectively, were amplified from cDNA and cloned into the multiple cloning site of the psiCHECK2 vector (Promega) by HiFi assembly (New England Biolabs). Control vectors were created by mutagenizing the predicted miRNA seed match in each of these vectors by PCR using 5' phosphorylated primers followed by ligation. For Taf7, 3'UTR fragments containing wild-type or mutant binding site for miR-21-5p was synthesized and cloned into the psiCHECK2 vector by HiFi assembly. The luciferase reporters were transfected into MEFs in triplicates. Luciferase activity was measured using the dual-luciferase reporter assay system (Promega) according to manufacturer's instructions 48 hrs post transfection.

Two-color fluorescent reporter assay

MEFs were engineered to stably express the reverse tetracycline-controlled transactivator (rtTA) using a lentiviral vector rtTA-N144 (Addgene: 66810) (Richner et al., 2015). The two-color fluorescent reporter pTRETightBI-RY-0 (Addgene: 31463), pTRETightBI-RY-1pf (Addgene: 31467) and pTRETightBI-RY-4 (Addgene: 31465) were transfected into the rtTA-expressing MEFs. 48 hrs after transfection, fluorescent signals were measured using flow cytometry. Signals were processed as described by Mukherji and colleagues (Mukherji et al., 2011). Mean and standard deviation of autofluorescence in eYFP and mCherry channels were obtained from untransfected cells. The mean autofluorescence plus twice the standard deviation was subtracted from each cell's eYFP and mCherry signals. Cells with eYFP signals lower than 0 were removed. The fluorescent signals were binned along the eYFP axis and mean mCherry signals were calculated in each bin.

RNA sequencing

Total RNAs from MEFs were extracted using TRIzol Reagent and subjected to DNase (Qiagen) treatment followed by RNeasy column clean-up (Qiagen). After quantification and quality control, 500ng of total RNA underwent poly(A) selection and TruSeq library preparation using the TruSeq Stranded mRNA LT Kit (Illumina) according to the manufacturer's instructions. Samples were barcoded and run on a HiSeq 2500 in a 50bp/50bp paired end run.

Reads were aligned to the standard mouse genome (mm10) using STAR v2.5.3a (Dobin et al., 2013). RNA reads aligned were counted at each gene locus. Expressed genes were subjected to differential gene expression analysis by DESeq2 v1.20.0 (Love et al., 2014).

Generation and analysis of P13 cortex HEAP libraries

Cortices were collected from two P13 *Ago2^{Halo/+}* mice. The HEAP and input control libraries were generated following the procedures described before.

Peak calling was run using the two HEAP libraries against the two matching input control libraries using the following parameters: count.threshold = 5, extend.slice = 10, bandwidth = 80, extend.peaks.in.genes = 150.

Analysis of public datasets

The CLEAR-CLIP (GSE73059) dataset was obtained from GEO and mapped to the standard mouse genome (mm10). For peak calling, the same parameters for the P13 cortex HEAP libraries were used. Peaks were identified by comparing the 12 CLEAR-CLIP libraries to the input control libraries generated for HEAP. Differential HEAP read count analysis in HEAP vs. input control was performed using DESeq2 v1.20.0.

The microarray dataset from CAD cell expressing miR-124 was obtained from (GSE8498) (Makeyev et al., 2007) using function getGEO() from GEOquery

v2.50.5 (Davis and Meltzer, 2007). Differential expression analysis was run using functions `lmFit()` and `eBayes()` from `limma` v3.38.3 (Ritchie et al., 2015).

CHAPTER 4: MIRNA TARGET IDENTIFICATION IN DEVELOPING EMBRYOS AND AUTOCHTHONOUS CANCER MODELS

Introduction

In this chapter I will present our results aimed at applying the newly developed HEAP method to map the miRNA-RNA networks during mouse embryonic development and tumorigenesis.

Embryogenesis is a tightly regulated process, characterized by coordinated cell growth, death, differentiation and movement. The precise control of gene expression networks and cell behaviors by miRNAs is crucial for proper morphogenesis and organogenesis. One essential component in the miRNA regulation network is the miR-17~92 cluster, which plays pleiotropic roles in mouse embryonic development. Deletion of miR-17~92 causes a spectrum of abnormalities affecting many parts of the embryo, including skeleton, heart, lung and blood (Han et al., 2015; Ventura et al., 2008). Despite the extensive phenotypic characterizations, molecular targets of this cluster of miRNAs remain elusive. Therefore, we applied the HEAP method in developing embryos and for the first time determined the direct targets of miR-17~92. The results of these experiments will be discussed in the first part of this chapter.

Tumorigenesis is often associated with de-regulation of the miRNA pathway, which involves copy number alterations of miRNA genes and aberrant transcription control of miRNA expression. For example, the miR-17~92 cluster is amplified in B-cell lymphomas (Ota et al., 2004), while frequent deletion of miR-15a/16-1 is associated with B-CLL (Calin et al., 2002). miR-17~92 is also a transcriptional tar-

get of the oncogene c-Myc, resulting in frequent overexpression of miR-17~92 in a variety of cancers (He et al., 2005; O'Donnell et al., 2005). Conversely, miR-34 is downstream of the p53 transcription program to promote apoptosis (Chang et al., 2007; Raver-Shapira et al., 2007). It remains unclear how the miRNA pathway works in concert with the oncogenic and tumor-suppressive programs to drive tumorigenesis. To explore the miRNA-RNA networks in primary autochthonous tumors, we focused on three genetic mouse models of two types of human cancers and combined them with the Halo-Ago2 knock-in strain we have developed. These results will be presented in the second half of this chapter.

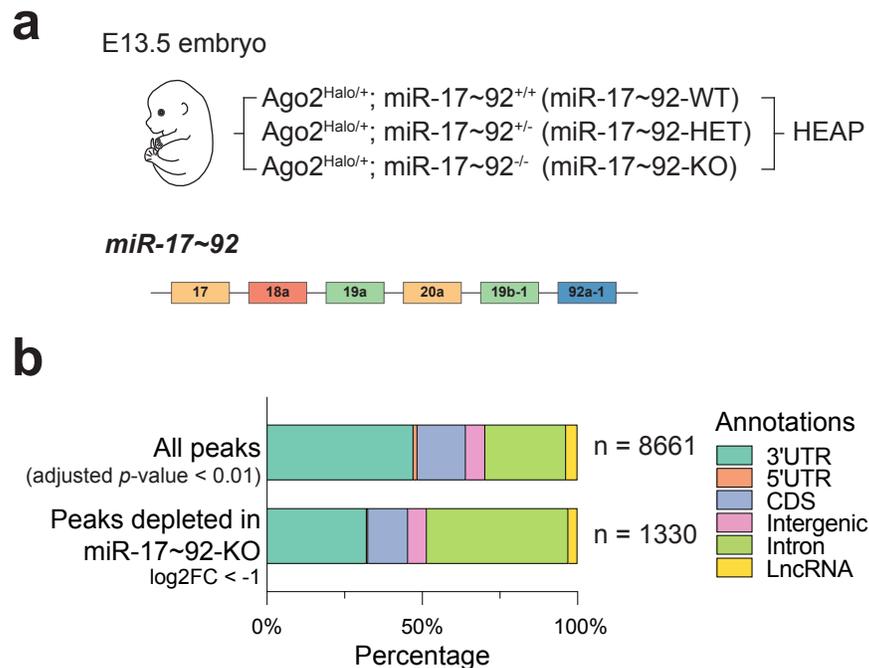


Figure 4.1: miRNA target identification in mid-gestation embryos

a) Outline of the HEAP experiments in E13.5 *Ago2^{Halo/+}* embryos wild-type, heterozygous or homozygous knockout for the miR-17~92 cluster. A schematic of the miR-17~92 cluster is shown at the bottom. miRNA members are color-coded based on their seed sequences. b) Bar plot showing the number and distribution across genomic annotations of peaks identified in the HEAP libraries from E13.5 embryos. “All peaks” indicates peaks identified across all genotypes. Distribution of peaks selectively depleted in the miR-17~92-KO embryo is shown in the lower panel.

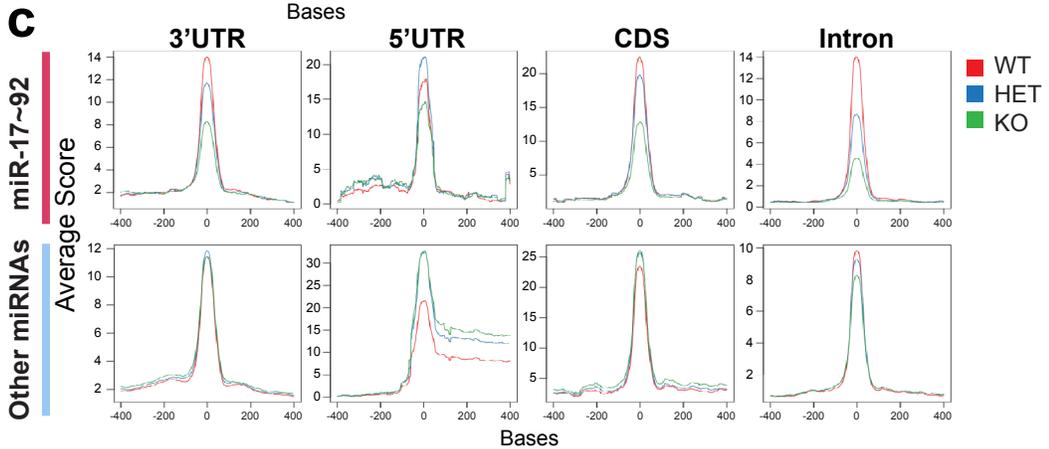
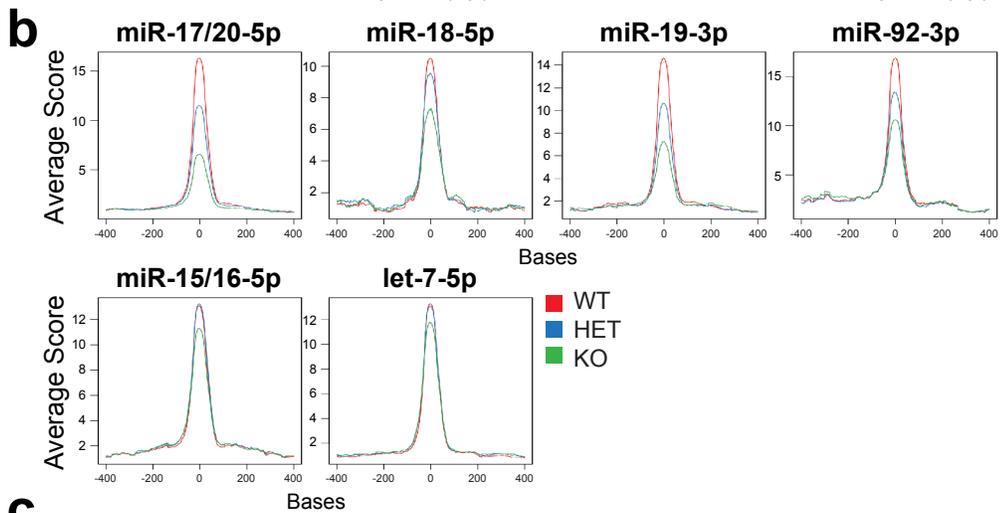
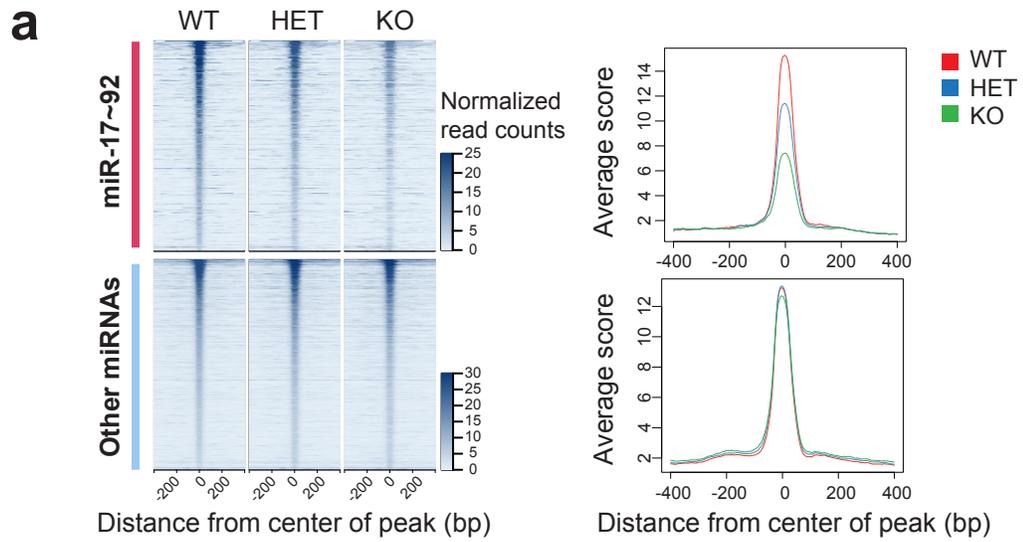
Characterization of miR-17~92 targets in developing embryos

To determine the direct targets of miR-17~92, we employed a mouse strain, in which the entire miR-17~92 cluster is deleted. The complete loss of miR-17~92 leads to perinatal lethality with full penetrance (Ventura et al., 2008). So, we crossed the *Ago2*^{Halo/+} mice ubiquitously expressing the Halo-Ago2 allele to mice harboring heterozygous deletion of the miR-17~92 cluster. We then generated HEAP libraries from *Ago2*^{Halo/+}; *miR-17~92*^{+/+} (miR-17~92-WT), *Ago2*^{Halo/+}; *miR-17~92*^{+/-} (miR-17~92-HET) and *Ago2*^{Halo/+}; *miR-17~92*^{-/-} (miR-17~92-KO) E13.5 embryos (**Figure 4.1a**). At an adjusted *p*-value cutoff of 0.01, HEAP identified a total of 8,661 peaks in these libraries, with a distribution across genomic annotations similar to that observed in mESCs (**Figure 4.1b**).

We then grouped the binding sites based on the occurrence of seed matches for members of the miR-17~92 cluster in their sequences. Importantly, the intensity of peaks overlapping with potential seed matches to miR-17~92 was greatly reduced—in a dose-dependent fashion—in the libraries generated from miR-17~92-HET and miR-17~92-KO embryos (**Figure 4.2a**). The dose-dependent reduction in peak signal was observed in peaks identified for each miRNA family in the miR-17~92 cluster, as well as in peaks located across genomic annotations

Figure 4.2: Dose-dependent loss of miR-17~92 targets in miR-17~92 knock-out embryos

a) Heatmap and histogram of peak signal in an 800 bp region surrounding HEAP peaks obtained from miR-17~92-WT, miR-17~92-HET and miR-17~92-KO E13.5 embryos. Peaks containing seed matches for the top 31 miRNA seed families ranked by abundance were chosen. Peaks with seed matches for miRNAs belonging to the miR-17~92 cluster are plotted in the upper panels, while the remaining peaks are plotted in the lower panels. **b)** Histogram of signal of peaks containing seed matches for miRNAs in the miR-17~92 cluster or for control miRNAs. **c)** Histogram of signal of peaks identified in different genomic regions. Peaks were grouped as in **a**).



(Figure 4.2b-c). In the miR-17~92-KO embryo, residual Halo-Ago2 binding signal was detected at many putative miR-17~92 binding sites. This could be explained by the presence of two additional paralogs to miR-17~92 (miR-106a~363 and miR-106b~25) that express similar miRNAs (Ventura et al., 2008) or targeting of the same sites by miRNAs from other seed families in the absence of miR-17~92.

By contrasting the miR-17~92-KO with the miR-17~92-HET and miR-17~92-WT libraries, we determined a class of targets that are dependent on miR-17~92

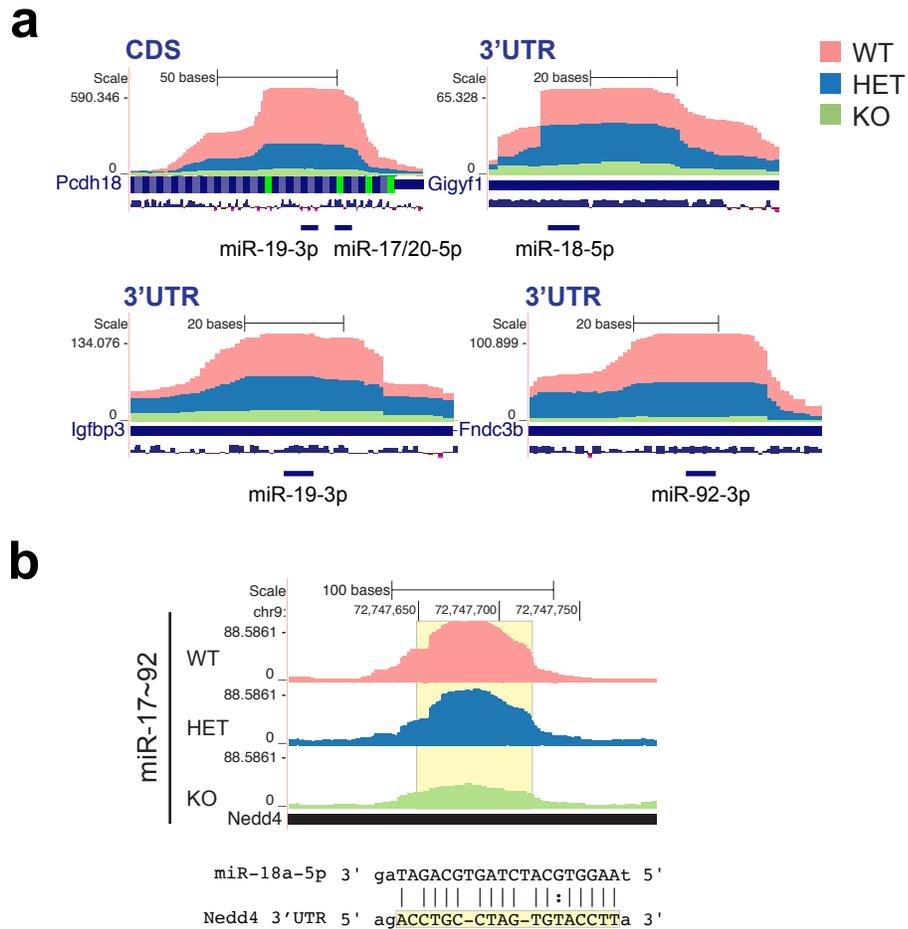


Figure 4.3: Representative miR-17~92-dependent binding sites

a) Genome browser views of representative canonical binding sites for miR-17/20-5p, miR-18-5p, miR-19-3p and miR-92-3p. Notice the potential cooperation between miR-17/20-5p and miR-19-3p in the 3'UTR of *Pcdh18*. **b)** Genome browser view of a non-canonical binding site for miR-18a-5p. Predicted pairing mode between miR-18a-5p and the 3'UTR of *Nedd4* is shown at the bottom.

expression (**Figure 4.1b**). Targets for the miR-17/20-5p, miR-18-5p, miR-19-3p and miR-92-3p seed families were visualized in UCSC genome browser (**Figure 4.3a**). In addition, we also observed non-canonical binding sites to miR-17~92 (**Figure 4.3b**).

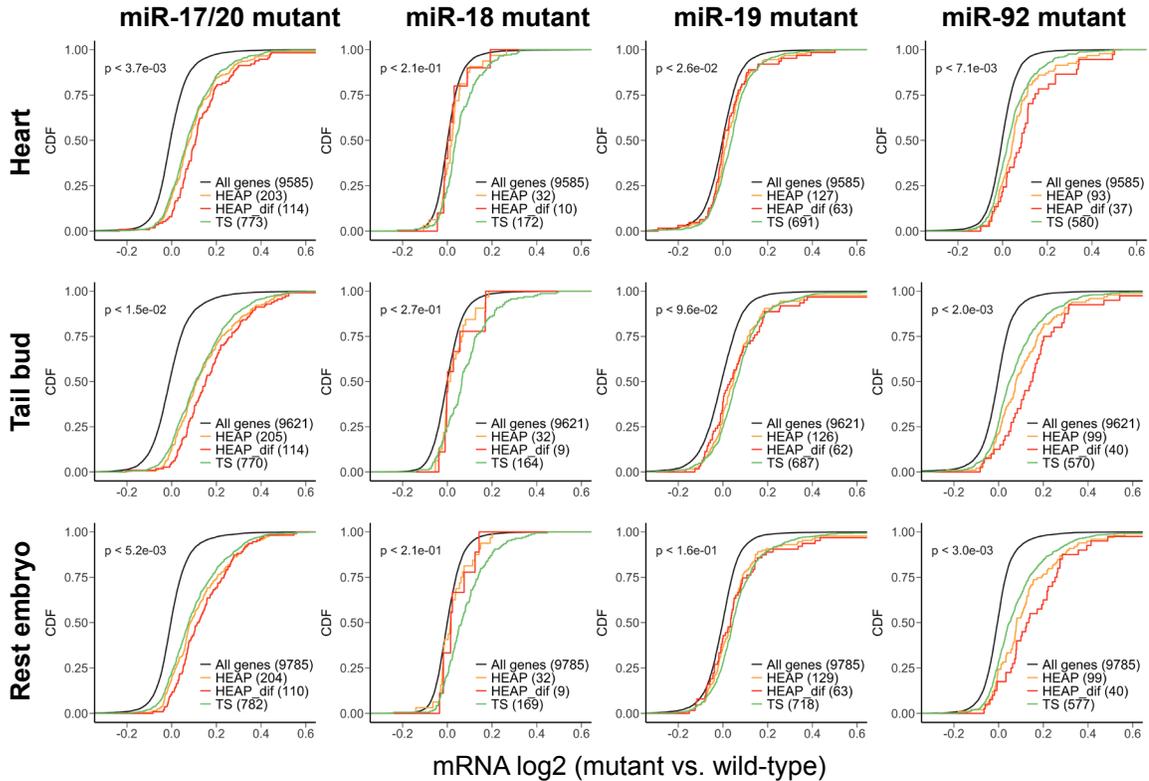


Figure 4.4: Preferential target mRNA repression associated with HEAP binding sites

CDF plots for all HEAP targets ("HEAP"), HEAP targets preferentially depleted in the miR-17~92-KO embryo ("HEAP dif") and targets predicted by TargetScan ("TS") for indicated miRNA seed families. The mRNA log₂ fold change for a particular miRNA (e.g. miR-17/20) was calculated in the miR-17~92 mutant E9.5 embryos by contrasting all conditions that are mutant for this miRNA (e.g. Δ17, Δ17,18, Δ17,18,92 and KO) to all other conditions (e.g. WT, Δ18, Δ19, Δ92). KO: embryos null for the entire miR-17~92 cluster; Δ17: embryos null for miR-17 and miR-20a; Δ18: embryos null for miR-18a; Δ19: embryos null for miR-19a and miR-19b-1, Δ92: embryos null for miR-92a-1; Δ17,18: embryos null for miR-17, miR-18a and miR-20a; Δ17,18,92: embryos null for miR-17, miR-18a, miR-20a and miR-92a-1. *P*-values were calculated between "HEAP dif" and "TS" using two-sided Kolmogorov–Smirnov tests.

To determine the activity of these binding sites in directing gene silencing, we re-analyzed an RNA-seq dataset previously generated in our laboratory from E9.5 embryos harboring an allelic series of miR-17~92 mutant alleles [(Han et al., 2015), GSE63813]. We demonstrated that HEAP targets containing seed matches for miR-17/20-5p, miR-19-3p and miR-92-3p mediated strong target repression (**Figure 4.4**). The effect was particularly evident when considering genes harboring HEAP peaks for miR-17/20-5p and miR-92-3p, whose signal intensity was reduced in the miR-17~92-KO embryo, confirming the importance of combining biochemical and genetic approaches to study miRNA function.

Interestingly, we identified a sizeable fraction of reproducible peaks (4%) mapping to non-coding RNAs. These included two previously uncharacterized miR-17~92-dependent peaks matching the miR-92-3p seed in the long non-coding RNA *Cyrano* (Kleaveland et al., 2018; Ulitsky et al., 2011) (**Figure 4.5a**). Supporting the hypothesis that these peaks are functionally relevant, we observed significant upregulation of *Cyrano* in mouse E9.5 embryos lacking miR-92a-1, but not in embryos harboring selective deletion of the other members of the cluster (Han et al., 2015) (**Figure 4.5b**). These results demonstrate the usefulness of the Halo-Ago2 mouse strain in facilitating the identification of miRNA targets *in vivo* and suggest that additional studies aimed at determining the functional consequences of loss of miR-92a-1 on *Cyrano* function may be warranted.

Characterization of miRNA targets in Bcan-Ntrk1 gliomas and cortices

Next, we chose a mouse model of glioma that we recently developed in our laboratory (Cook et al., 2017) to characterize miRNA regulation signatures during oncogenic transformation. In this model, gliomas are driven by the *Bcan-Ntrk1* gene fusion. *p53^{fl/fl}* mice are injected intracranially with a mixture of two recombi-

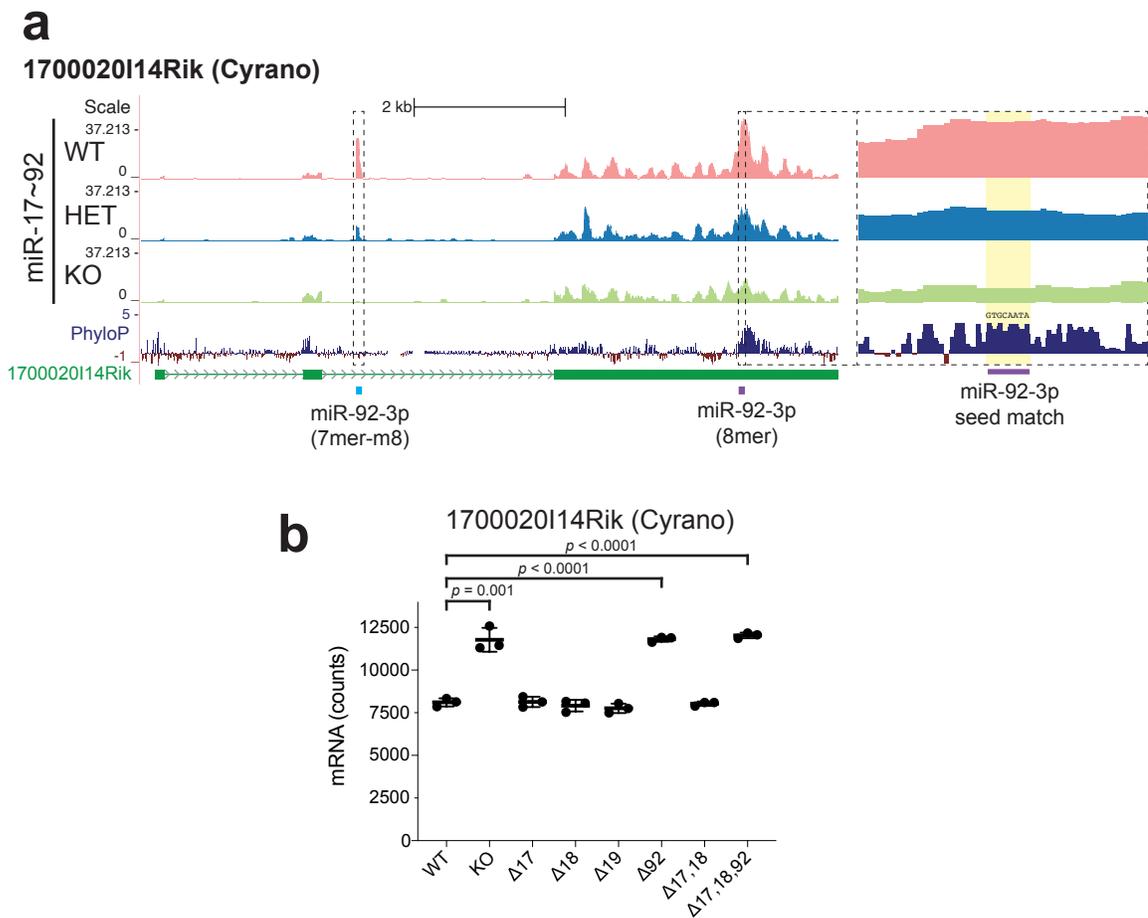


Figure 4.5: Regulation of a lncRNA by miR-92-3p

a) Genome browser view of the miR-92a-1-dependent miRNA binding sites detected in the long non-coding RNA *1700020I14Rik* (*Cyrano*). PhyloP: placental mammal basewise conservation by PhyloP. Notice the highly conserved 8mer seed match for miR-92-3p under the second peak. **b)** RNA expression of *Cyrano* in the heart of E9.5 embryos harboring an allelic series of miR-17~92 mutant alleles. KO: embryos null for the entire miR-17~92 cluster; Δ17: embryos null for miR-17 and miR-20a; Δ18: embryos null for miR-18a; Δ19: embryos null for miR-19a and miR-19b-1, Δ92: embryos null for miR-92a-1; Δ17,18: embryos null for miR-17, miR-18a and miR-20a; Δ17,18,92: embryos null for miR-17, miR-18a, miR-20a and miR-92a-1. Notice that *Cyrano* is only up-regulated in mutants in which miR-92a-1 is deleted. *P*-value: unpaired t-test. Data are represented as Mean ± SD.

nant adenoviruses. The first expresses Cas9 and two sgRNAs (Ad-BN) designed to induce the *Bcan-Ntrk1* rearrangement, an intra-chromosomal deletion resulting in the fusion between the N-terminal portion of Bcan and the kinase domain of Ntrk1.

The second adenovirus expresses the Cre recombinase (Ad-Cre) to achieve concomitant deletion of p53 and allow glioma formation. By performing this procedure in 4~6-week-old *p53^{fl/fl}; Ago2^{Halo-LSL/+}* mice, we produced Bcan-Ntrk1 driven gliomas stably expressing the endogenous Halo-Ago2 allele. The intracranial adenovirus administration was performed by Peter J. Cook.

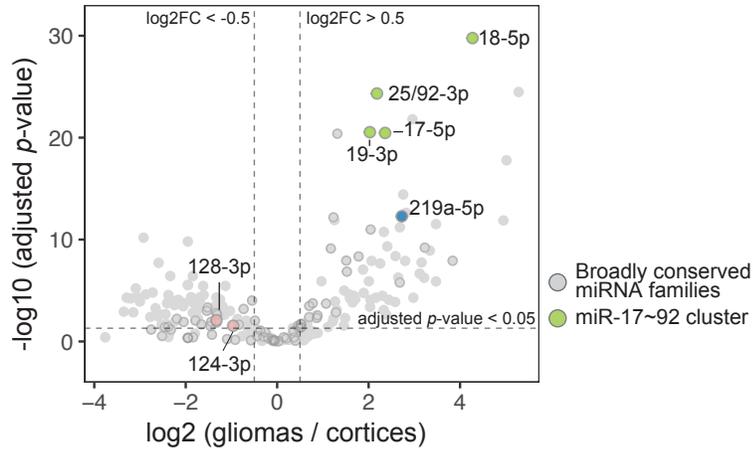


Figure 4.6: Differential expression of miRNAs between gliomas and normal cortices

Volcano plot of global changes in miRNA family expression between normal cortices and Bcan-Ntrk1-driven gliomas, as determined by HEAP. Broadly conserved families are highlighted in circles and a few miRNA families of interest are colored and annotated.

We generated HEAP libraries from three independent Bcan-Ntrk1 gliomas and from the normal cortices of three age-matched *Ago2^{Halo/+}* mice. Quantification of miRNA abundance in HEAP miRNA libraries revealed drastic differences between the two tissues, with 77 miRNA seed families (26 broadly conserved) being significantly upregulated in gliomas, and 77 families (18 broadly conserved) downregulated (adjusted *p*-value < 0.05, absolute log₂FC > 0.5, **Figure 4.6**). Of note, the significantly downregulated families include miR-124-3p and miR-128-3p, two miRNA families that are highly expressed in the cortex of mice (Bak et al., 2008;

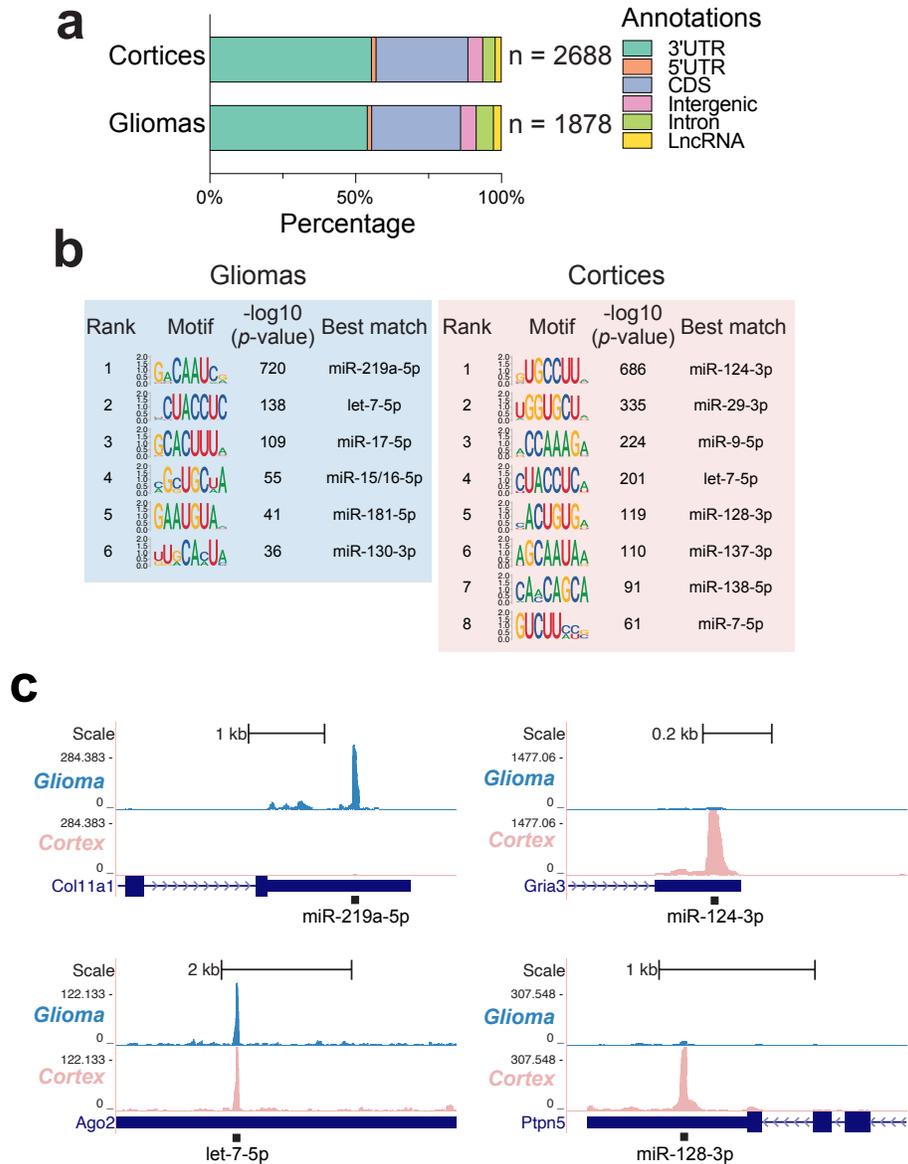


Figure 4.7: Distinct miRNA targeting landscapes between gliomas and normal cortices

a) Total number and distribution across genomic annotations of peaks identified in the cortex and glioma HEAP libraries at adjusted p -value < 0.05 . **b)** Top differentially enriched 8-mers in glioma and cortex HEAP peaks (peak selection cutoff: adjusted p -value < 0.05 ; absolute $\log_2(\text{gliomas} / \text{cortices}) > 0.5$) by the HOMER *de novo* motif discovery algorithm. miRNA families whose seed sequences are complementary to these motifs are annotated. **c)** Representative peaks containing seed matches for miRNA families that are differentially present in the two tissues (miR-219a-5p, miR-124-3p, miR-128-3p) and for a miRNAs family whose targets are found in both conditions (let-7-5p).

Landgraf et al., 2007b) and functionally important in the mouse central nervous system as suggested by genetic loss-of-function studies (Sanuki et al., 2011; Tan et al., 2013). Additionally, members of the oncogenic miRNA cluster miR-17~92 (He et al., 2005; Ota et al., 2004) were among the most strongly upregulated miRNAs in gliomas, suggesting the possibility that these miRNAs are functionally relevant in gliomagenesis.

Using an adjusted *p*-value cutoff 0.05, we identified 1,878 Halo-Ago2 binding sites in tumors and 2,688 sites in normal cortices, with an overlap of 1,335 sites. Peaks distribution across genomic annotations was similar between the two tissues, with the majority of peaks mapping to 3'UTRs (**Figure 4.7a**). Analysis of seed matches under the peaks revealed marked differences between normal and neoplastic brains. Motifs complementary to the seeds of miR-219a-5p, miR-17-5p, miR-15/16-5p, miR-181-5p and miR-130-3p were preferentially enriched in peaks identified in gliomas, while motifs complementary to the seeds of miR-124-3p, miR-29-3p, miR-9-5p, miR-128-3p, miR-137-3p, miR-138-5p and miR-7-5p were preferentially enriched in peaks from normal cortices (adjusted *p*-value < 0.1, absolute log₂FC (gliomas vs. cortices) > 0.5). Targets for the let-7-5p family of miRNAs were also abundant, but not differentially represented between the normal brain and tumors (**Figure 4.7b**). The differential enrichment of motifs was corroborated by visualizing these peaks on the UCSC genome browser (**Figure 4.7c**).

To understand the enrichment for specific seed matches observed in the two conditions, we computed differential peak signal changes for targets of all broadly conserved miRNA families. The medians of the peak signal changes were plotted against the abundance changes for members in the corresponding miRNA families. As a result, the differences observed in peak intensities were largely explained by the differences in miRNA levels (**Figure 4.8a**). The size of the circles is

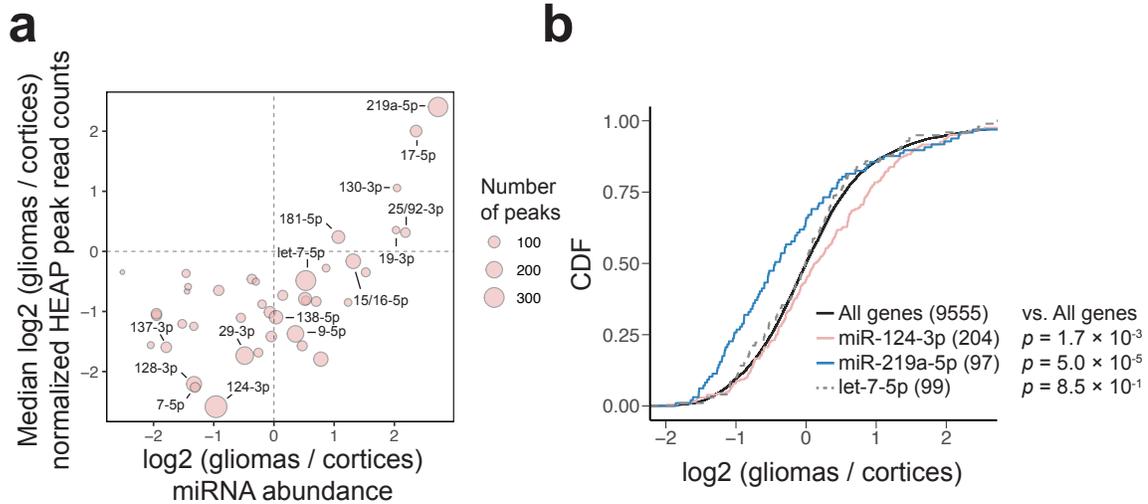


Figure 4.8: Enrichment of miRNA seed matches correlates with miRNA abundance and causes differential repression of corresponding targets

a) Changes in peak intensity correlate with changes in miRNA abundance. The area of each circle is proportional to the number of targets of each miRNA seed family as identified by HEAP. Only broadly conserved miRNA families with more than 40 HEAP targets are shown. **b)** CDF plot for targets of miR-124-3p, miR-219a-5p and let-7-5p identified by HEAP. mRNA expression was estimated using read counts in input control libraries. The mRNA log₂ fold change was calculated as gliomas vs. cortices. *P*-value: two-sided Kolmogorov–Smirnov test.

proportional to the number of binding sites identified for the corresponding miRNA families. The strongest target intensity differences were observed for the miR-219a-5p and miR-124-3p families, with their targets being highly enriched in glioma and cortex, respectively. The differences in target signals were translated into differential gene regulation, as demonstrated by a statistically significant repression of miR-219a-5p targets in gliomas and of miR-124-3p targets in the normal cortices (**Figure 4.8b**).

Among all miRNA families, the miR-219a-5p family had the highest number of targets in gliomas (300 out of 1,878 peaks contained 6mer, 7mer or 8mer seed matches to miR-219a-5p; **Figure 4.9a**). miR-219a-5p has been reported to regulate oligodendrocyte (OL) differentiation and myelination in mice via targeting important regulators of oligodendrocyte progenitor cell (OPC) maintenance (Dugas et

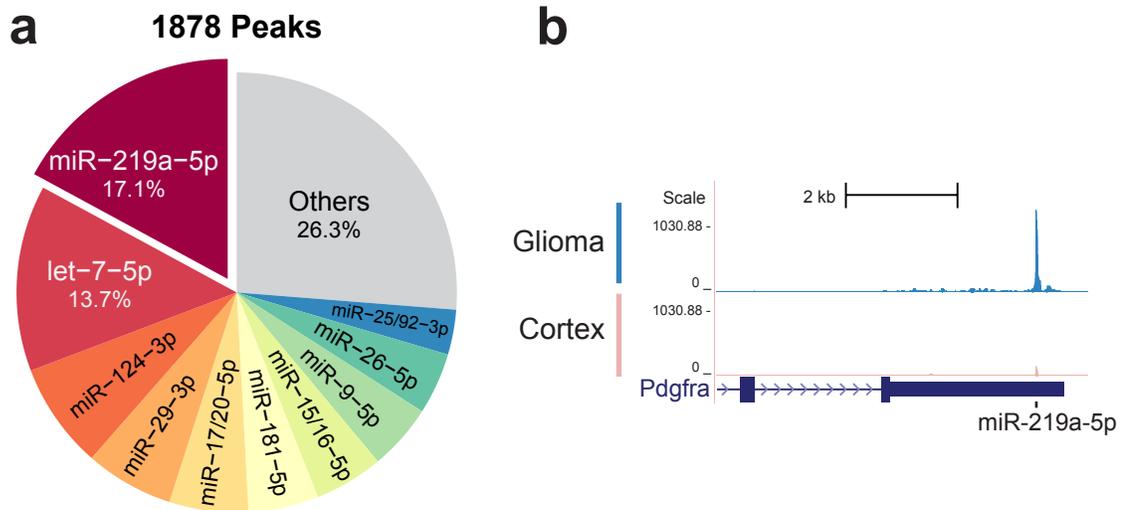


Figure 4.9: Enrichment of miR-219a-5p binding sites in gliomas

a) Pie chart showing the distribution of Halo-Ago2 binding sites identified in gliomas among miRNA seed families. b) Genome browser view of a miR-219a-5p binding site in *Pdgfra*.

al., 2010; Emery, 2010; Fan et al., 2017; Wang et al., 2017; Zhao et al., 2010). Interestingly, we observed a strong interaction between miR-219a-5p and *Pdgfra*, a characteristic marker of OPCs and a key player in gliomagenesis (**Figure 4.9b**).

In summary, the characterization of miRNA binding sites in the *Bcan-Ntrk1* driven gliomas and normal cortices delineated the miRNA targeting landscapes and highlighted notable differences between the two tissues. By performing motif enrichment analysis, we determined key miRNA targeting signatures that distinguished tumors from their tissues of origins. Finally, a positive correlation between target signal and miRNA abundance was observed in both contexts, suggesting that it is likely the expression of miRNAs that shapes the miRNA regulation networks.

Characterization of miRNA targets in lung adenocarcinomas

The clear context-dependent miRNA expression and targeting events re-

vealed in the previous section suggested further analysis on determining whether oncogenic driver events could potentially shape the miRNA targeting landscapes in cancer cells. To address this question, we mapped miRNA-RNA interactions in two murine models of non-small cell lung cancer (NSCLC): the Cre recombinase-mediated $KRas^{LSL-G12D/+}; Trp53^{fl/fl}$ (KP) model (Jackson et al., 2001) and a CRISPR-Cas9 induced model driven by a chromosomal inversion resulting in the formation of the *Eml4-Alk* (EA) gene fusion (Maddalo et al., 2014). These two mouse models recapitulate two types of NSCLC observed in humans and differ not only in the initiating genetic lesions but also in the modality with which tumor formation is induced.

We generated HEAP libraries from $Ago2^{Halo-LSL/+}$ mice bearing primary KP (N = 2) and EA (N = 3) tumors. Tumor-specific expression of the Halo-Ago2 allele was induced at the time of tumor initiation by intratracheal delivery of Ad-Cre, alone for

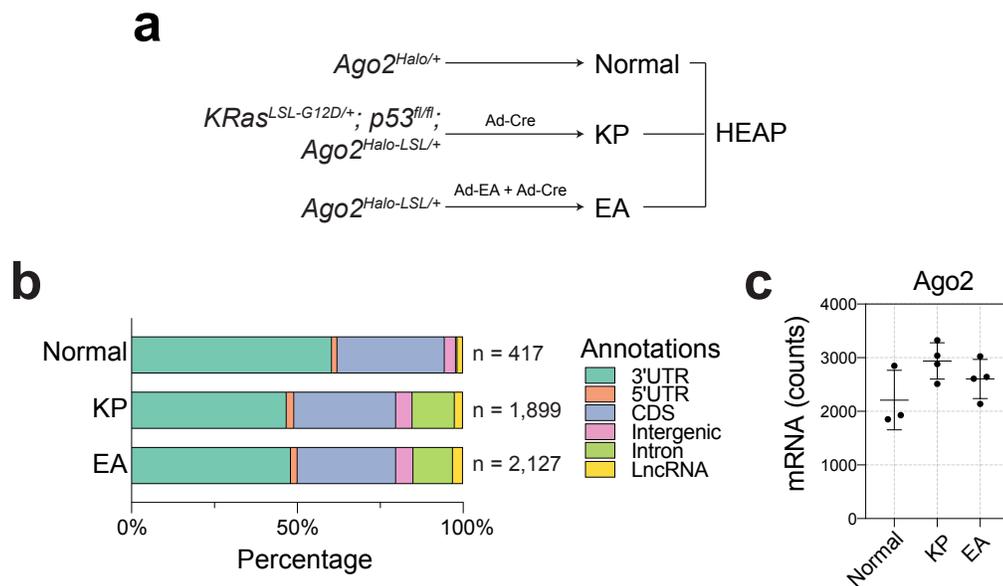


Figure 4.10: miRNA target identification in lung adenocarcinomas and normal lungs

a) Schematic of the experimental design. **b)** Total number and distribution across genomic annotations of peaks identified in normal lungs (two replicates) and in the KP and EA lung adenocarcinomas (two and three replicates, respectively) at adjusted p -value < 0.05. **c)** Normalized counts for *Ago2* mRNA in normal lungs and in KP and EA lung tumors. Data are represented as Mean ± SD.

the KP model or in combination with recombinant adenoviruses expressing Cas9 and the two sgRNAs necessary to induce the *Eml4-Alk* rearrangement in the EA model (Ad-EA). In parallel, we also generated HEAP libraries from the lungs of two *Ago2*^{Halo/+} mice (**Figure 4.10a**).

The tumor libraries produced 1,899 peaks for the KP tumors and 2,127 peaks for the EA tumors. In contrast, only 417 peaks were identified in normal lungs (**Figure 4.10b**). This difference could not be attributed to differences in sequencing depth or Halo-Ago2 expression levels in normal lungs vs. tumors (**Figure 4.10c**). Rather, it may reflect reduced levels of fully assembled miRISC in the normal lung compared to lung tumors [(La Rocca et al., 2015) and La Rocca et al., manuscript in preparation].

Surprisingly, a direct comparison of the peaks identified in KP and EA tumors revealed strong similarity between the two tumor types (**Figure 4.11a**), suggesting that the miRNA targeting landscape is largely independent from the cancer initiation events in these two NSCLC models. Unbiased k-mer frequency analysis visualized as motif enrichment identified distinct miRNA seed matches enriched in peaks in normal lung and tumors. Binding sites for let-7-5p, miR-29-3p and miR-30-5p were strongly enriched in both tissues, while seed matches for several miRNAs implicated in tumorigenesis and metastasis, such as miR-200bc-3p (Davalos et al., 2012; Gibbons et al., 2009; Gregory et al., 2008; Sato et al., 2017; Si et al., 2017), miR-31-5p (Edmonds et al., 2016), miR-17-5p (He et al., 2005; Ota et al., 2004) and miR-25/92-3p (Ota et al., 2004) were dominant in the tumor libraries (**Figure 4.11b**). In human lung adenocarcinomas, miR-200 levels negatively correlate with tumor metastatic potential, at least in part because this miRNA can potentially suppress epithelial-to-mesenchymal transition (EMT) (Davalos et al., 2012; Gibbons et al., 2009; Si et al., 2017). In agreement with this model, we observed a

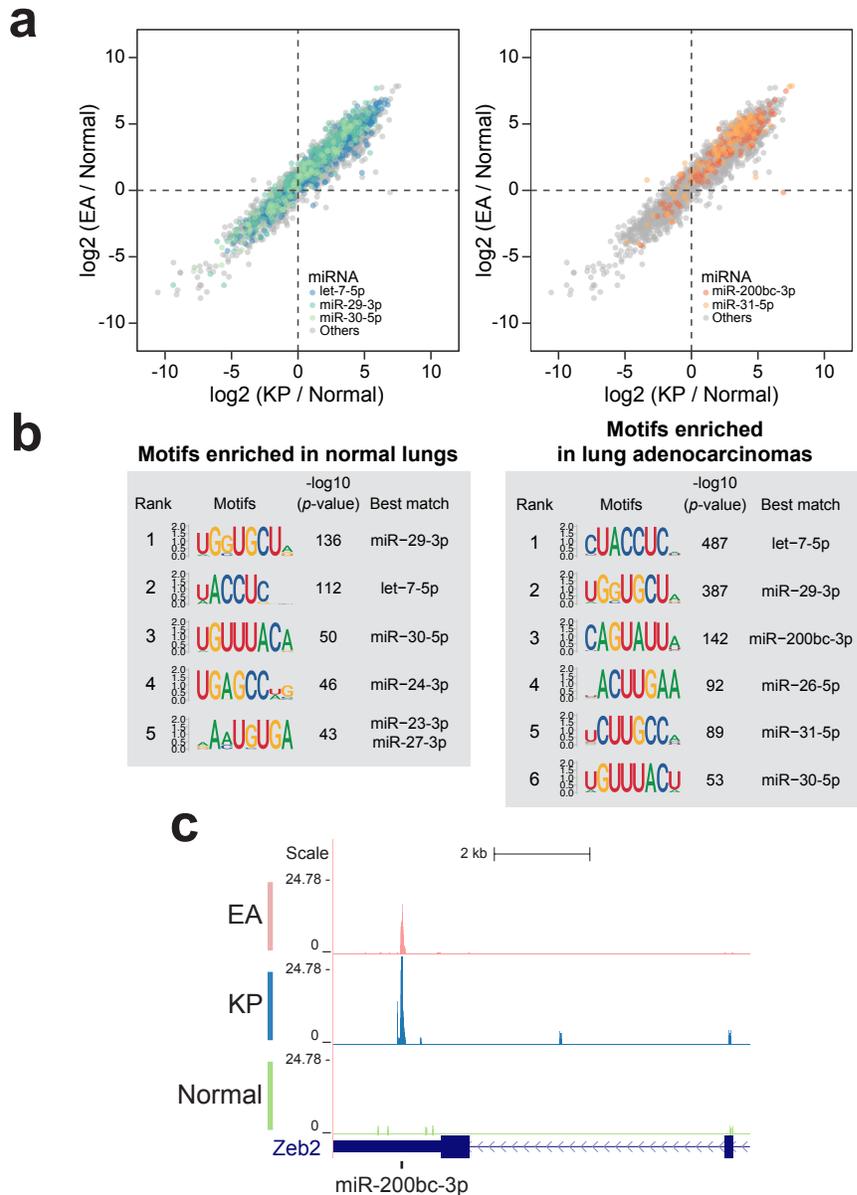


Figure 4.11: Similar miRNA targeting landscapes between the two NSCLC models

a) Scatter plots to compare peak intensity changes in EA vs. normal (y-axis) and KP vs. normal (x-axis). Peaks with seed matches for indicated miRNA families are colored. Left: highlighted are peaks containing seed matches for representative miRNA families whose targets were found in both normal lungs and lung adenocarcinomas. Right: highlighted are peaks with seed matches for miRNA families whose targets were preferentially found in lung adenocarcinomas. **b**) Sequence logos of the most enriched 8-mers as determined by the HOMER *de novo* motif discovery algorithm under HEAP peaks identified in normal lungs (left) and in lung adenocarcinomas (right). miRNA families whose seeds are complementary to these motifs are annotated. **c**) Genome browser view of the miR-200bc-3p binding site in the 3'UTR of *Zeb2*.

strong miR-200bc-3p binding site in the 3'UTR of *Zeb2*, a master regulator of EMT (**Figure 4.11c**).

To further validate the functional significance of these miRNA-RNA interactions in lung cancer, we compared the transcriptome of mouse KP cancer cells expressing either T6B-YFP (T6B^{WT}-YFP) or a mutant version (T6B^{MUT}-YFP) that cannot bind Ago proteins and is therefore inactive [LaRocca et al., manuscript in preparation]. As shown in **Figure 4.12**, genes harboring peaks identified by HEAP were preferentially de-repressed upon disruption of the miRISC, further confirming the ability of the HEAP method to identify functional miRNA-RNA interactions *in vivo*. The generation of KP cell lines expressing the T6B-YFP fusion and the RNA-seq experiment were performed by Minsi Zhang, a postdoctoral fellow in our laboratory.

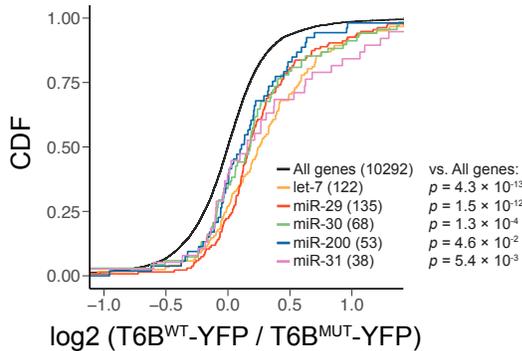


Figure 4.12: Preferential gene repression associated with HEAP 3'UTR binding sites in KP cells

CDF plot of mRNA expression changes induced by T6B-YFP expression in murine KP cells (T6B^{WT}-YFP / T6B^{MUT}-YFP). Targets identified by HEAP for the indicated miRNA families were compared to background (“all genes”). *P*-value: two-sided Kolmogorov–Smirnov test.

Collectively, the miRNA binding site analysis in the two mouse models of human lung adenocarcinomas suggested highly similar miRNA regulation signa-

tures. A parallel RNA-seq experiment in the KP and EA tumors revealed very similar gene expression profiles (data not shown). As EA and KP tumors are thought to originate from the same cells in mice—the alveolar type 2 cells in the lung (Sutherland et al., 2014), it might be the cell type and the overall transcription programs that determined the miRNA expression and target profiles

Discussion

Identification of miR-17~92 targets in the whole embryo is only the first step towards functional dissection of this miRNA cluster. miR-17~92 is ubiquitously expressed across lineages during gastrulation and organogenesis (Pijuan-Sala et al., 2019). Its expression can be detected as early as blastocyst stage (Foshay and Gallicano, 2009). Precise spatial and temporal control of miR-17~92 expression has also been reported in the B-cell compartment (Ventura et al., 2008). The HEAP experiment only took a snapshot of miRNA regulations in many concurrent biological processes. Therefore, it is computationally challenging to decipher the roles of miR-17~92 only based on targets captured in this experiment. A more focused target discovery in a specific tissue is recommended. Furthermore, miRNAs are also hypothesized to regulate cell fate switch by providing robustness to the transcription programs (Ebert and Sharp, 2012). It would be technically difficult to capture the targets if the key interaction events are only transiently present during embryonic development.

The comparison of miRNA targeting events between gliomas and cortices highlighted the functional importance of miR-219a-5p family in gliomas. The enrichment of miR-219a-5p targeting signature in gliomas reflects the cell of origin of these tumors – glial cells. miR-219a-5p is the most strongly induced miRNA during the differentiation of oligodendrocytes (OL), a subtype of glial cells. miR-219a-5p

is required for OPC differentiation (Dugas et al., 2010; Zhao et al., 2010). It is possible that the gliomas characterized in this study transcriptionally resemble differentiating OPCs. Using HEAP, we have also corroborated an important interaction event between miR-219a-5p and *Pdgfra* during OPC differentiation as suggested by Dugas and colleagues (Dugas et al., 2010) (**Figure 4.9b**). In line with this speculation, a recent single-cell RNA-seq study suggests that glioblastoma cells recapitulate cells in several neurodevelopmental trajectories of glial cells in developing brains, including cells in the OL lineage (Couturier et al., 2020).

Altogether, we have characterized miRNA binding sites in three models of two types of human cancers. Shared by these results is the enrichment of miR-17~92 targets in all models. miR-17~92 is also differentially expressed in glioma cells, which is possibly driven by the upregulation of *c-Myc* in gliomas (data not shown). Whether the activation of miR-17~92 plays pro-tumorigenic roles remains to be examined.

miR-29-3p and let-7-5p are two ubiquitously expressed miRNA families, whose targets are highly enriched in both normal adult tissues and tumors. In all cases, large numbers of targets for miR-29-3p and let-7-5p were identified, suggestive of a general requirement of these miRNA families for homeostasis maintenance and pathogenesis.

Finally, the HEAP experiments in the primary tumors nicely showcased target purification from live tissues in a cell-type-specific manner using the conditional Halo-Ago2 knock-in mice. In all cases, the Halo-Ago2 fusion was turned on irreversibly in the tumor cell of origin concomitantly with the oncogenic programs necessary for driving tumor formation. For the generation of HEAP libraries, tumors were resected from adjacent normal tissues under a dissecting microscope and directly subjected to crosslinking, lysis and target purification. The use of the

Halo-Ago2 knock-in mice greatly simplifies the *in vivo* target discovery of miRNAs.

Method

Animal models

miR-17~92^{-/-} (Ventura et al., 2008), *Trp53^{fl/fl}* (Marino et al., 2000) and *KRas^{LSL-G12D/+}* (Jackson et al., 2001) mice were used in this study. For the generation of E13.5 embryos, 6~10-week-old females were sacrificed at embryonic day 13.5. To generate P13 cortex HEAP libraries, cortices were harvested from 13-day-old *Ago2^{Halo/+}* mice. For the generation of gliomas, 4~6-week-old *Ago2^{Halo-LSL/+}; Trp53^{fl/fl}* mice were infected with recombinant adenoviruses and tumors were harvested approximately 80 days after injection. Normal cortices were harvested from age-matched *Ago2^{Halo/+}* mice. For the generation of lung adenocarcinomas, 10~12-week-old *Ago2^{Halo-LSL/+}* (EA model) and *Ago2^{Halo-LSL/+}; KRas^{LSL-G12D/+}; Trp53^{fl/fl}* (KP model) mice were infected with recombinant adenoviruses and tumors were harvested 3 months after infection. Normal lungs were obtained from age-matched *Ago2^{Halo/+}* mice.

All studies and procedures were approved by the Memorial Sloan Kettering Cancer Center Institutional Animal Care and Use Committee.

Recombinant adenovirus delivery

Recombinant adenoviruses used for inducing chromosomal rearrangements (Ad-BN, Ad-EA) (Cook et al., 2017; Maddalo et al., 2014) and Ad-Cre were purchased from ViraQuest.

For the generation of Bcan-Ntrk1-driven gliomas, a 1:1 mixture of Ad-BN and Ad-Cre, in total $\sim 3 \times 10^9$ infectious particles, was administered to *Ago2^{Halo-LSL/+}; Trp53^{fl/fl}* mice (4~6 weeks old), via stereotactic intracranial injection as described in Cook et al., 2017. Gliomas were harvested approximately 80 days after injection, when mice became symptomatic.

For the generation of Eml4-Alk-driven lung adenocarcinomas, 10~12-week-old *Ago2^{Halo-LSL/+}* mice were intratracheally infected with a 1:1 mixture of Ad-EA and Ad-Cre (in total $\sim 6 \times 10^{10}$ infectious particles). To generate *KRas^{G12D}; Trp53^{-/-}* lung tumors, 10~12-week-old *Ago2^{Halo-LSL/+}; KRas^{LSL-G12D/+}; Trp53^{fl/fl}* mice were intratracheally infected with Ad-Cre ($\sim 2.5 \times 10^7$ PFU). Lung tumors were harvested approximately 3 months after infection.

RNA sequencing

Total RNAs from lung adenocarcinomas and normal lung tissues were extracted using TRIzol Reagent (Invitrogen) and subjected to DNase (QIAGEN) treatment followed by RNeasy column clean-up (QIAGEN). After quantification and quality control, 500ng of total RNA underwent poly(A) selection and TruSeq library preparation using the TruSeq Stranded mRNA LT Kit (Illumina) according to the manufacturer's instructions. Samples were barcoded and run on a HiSeq 2500 in a 50bp/50bp paired end run.

Total RNAs of T6B-YFP-expressing KP cells were isolated using TRIzol Reagent and subjected to DNase treatment and isopropanol re-precipitation. After quantification and quality control, 1 ug of total RNA underwent ribosomal depletion and library preparation using the TruSeq Stranded Total RNA LT Kit (Illumina). Samples were run on a HiSeq 4000 in a 50bp/50bp paired end run.

Reads were aligned to the standard mouse genome (mm10) using Hisat2 (v0.1.6-beta) (Kim et al., 2019) or STAR v2.5.3a (Dobin et al., 2013). RNA reads aligned were counted at each gene locus. Expressed genes were subjected to differential gene expression analysis by DESeq2 v1.20.0 (Dobin et al., 2013).

Analysis of public datasets

RNA-seq data generated from E9.5 miR-17~92 mutant embryos were obtained from the authors and are available in GEO (GSE63813) (Han et al., 2015).

In this study, gene expression was profiled in triplicates in heart, mesoderm and all remaining tissues of wild-type (WT) embryos and embryos null for miR-17 and miR-20a ($\Delta 17$), null for miR-18a ($\Delta 18$), null for miR-19a and miR-19b-1 ($\Delta 19$), and null miR-92a-1 ($\Delta 92$), null for miR-17, miR-18a and miR-20a ($\Delta 17,18$), null for miR-17, miR-18a, miR-20a and miR-92a-1 ($\Delta 17,18,92$), and null for the entire cluster (KO). Embryos were of different genders. The data was aligned using HISAT v0.1.6-beta. In each tissue, differential gene expression analysis was performed using DESeq2 v1.6.3 using multi-factorial model “~ d17 + d18 + d19 + d92 + gender”, where factor “d17” encoded for conditions that were $\Delta 17$, factor “d18” encoded for conditions that were $\Delta 18$, etc., and factor “gender” encoded for the genders of the embryos. This allowed us to estimate the log₂FC of expression associated with each individual miRNA family in each tissue when accounting for contribution from other miRNA families and the gender.

Peak calling

For embryos, peak calling was run using HEAP in one wildtype (miR-17~92-WT), two heterozygous (miR-17~92-HET) and one homozygous knockout (miR-17~92-KO) embryo against the four matching input control libraries using the following parameters: count.threshold = 5, extend.slice = 10, bandwidth = 80, extend.peaks.in.genes = 150. Then differential HEAP read count analysis was performed using DESeq2 v1.22.1 in miR-17~92-KO against miR-17~92-WT and miR-17~92-HET libraries to determine miR-17~92-dependent peaks.

For gliomas and cortices in adult mice, three HEAP libraries from each context were generated. Before peak calling, size factors Y of the six HEAP libraries were estimated using the byte sizes of corresponding BAM files. Then, BAM files for two glioma replicates and three cortex replicates were downsampled to similar sizes to the smallest glioma replicate using samtools v1.3.1 (Li et al., 2009) with

scaling factors $X = 1/Y$. Peak calling was run using the six scaled HEAP libraries against the six matching input control libraries, using the same parameters as for embryos, to identify the set of putative peaks. As usual, only peaks of size > 20 nt and with $\log_2FC > 0$ in HEAP vs. input control were used in downstream analysis. Furthermore, only peaks with average normalized read count > 10 in the three glioma replicates or in the three cortex replicates were selected. To identify significant peaks in gliomas, DESeq2 v1.20.0 for read counts in these selected peaks was run using the three glioma replicates against the three matching input control replicates. To identify significant peaks in cortices, DESeq2 for read counts in these selected peaks was run using the three cortex replicates against the three matching input control replicates. Differential HEAP read counts analysis between gliomas and cortices was run in peaks with adjusted p -value < 0.05 (in HEAP vs. input control).

For lung tumors, peak calling was run using two HEAP libraries generated from normal lungs, two HEAP libraries from KP tumors and three HEAP libraries from EA tumors against seven matching input control libraries, using the same parameters as for embryos. Peaks of size > 20 nt and with $\log_2FC > 0$ in HEAP vs. input control were used in downstream analysis. Furthermore, only peaks with average normalized read count > 10 in the two normal lung replicates, in the two KP tumor replicates or in the three EA tumor replicates were selected. To identify significant peaks in each tumor type, DESeq2 v1.20.0 for read counts in these selected peaks was run using the tumor replicates against their matching input control replicates. To identify significant peaks in normal lungs, DESeq2 for read counts in the selected peaks was run using the two normal lung replicates against the two matching input control replicates. To compare peak intensities between KP and EA tumors, DESeq2 for read counts in peaks with adjusted p -value < 0.05

(in HEAP vs. input control) was run using the three EA tumor replicates against the two KP tumor replicates. Since peak intensities in EA and KP highly correlate with each other, the five tumor replicates were grouped and used for downstream analysis. To compare peak signals between tumors and normal lungs, differential HEAP read count analysis was performed in peaks with adjusted p -value < 0.05 (in HEAP vs. input control) between the five tumor replicates and the two normal lung replicates.

mRNA abundance estimates in the input control libraries

The input control libraries generated from gliomas and cortices were used to estimate mRNA abundance. Reads were counted at each gene locus using featureCounts v1.6.3 (Liao et al., 2014) with GENCODE (vM22) primary annotation. Differential gene expression analysis was performed using DESeq2 v1.20.0.

HOMER de novo motif discovery

In glioma and cortex libraries, the top 50 7-mers found by unbiased motif enrichment analysis in each context were mapped to corresponding peak set and subjected to HOMER *de novo* motif discovery as described before. For normal lung and lung tumor libraries, the top 70 7-mers from each context were used.

HEAP coverage analysis

To measure HEAP coverage of various peak sets in embryo libraries, peaks were first assigned to miRNA seed families by searching for the corresponding 7mer and 8mer seed matches in peak sequences. All miRNA seed families were ranked by abundance measured in miR-17~92-WT embryo. Peaks containing seed matches for the top 31 miRNA families were chosen. Score matrices of 800 bp windows surrounding these peaks were generated from size-factor-corrected bigWigs using the ScoreMatrixList() function from the genomation package v1.14.0 (Akalin et al., 2015). Histograms of average score were produced using the func-

tion `plotMeta()`. Heatmaps were generated using the `multiHeatMatrix()` function and extreme values were removed before plotting using the `winsorize` parameter with values `c(0,98)`.

CHAPTER 5: CONCLUSION AND PERSPECTIVE

Conclusions

In this thesis, I have reported the development of the HEAP method and demonstrated its ability to identify miRNA-RNA interaction sites in cells, developing embryos, normal adult tissues, and in primary autochthonous tumors. By mapping miRNA binding sites in mouse embryos lacking the miR-17~92 cluster, we identified direct targets of the miRNAs encoded by this cluster, including a long noncoding RNA that had not been previously reported to be regulated by this cluster. The HEAP method also allowed us to identify miRNA targets in primary autochthonous cancers in mice and in their tissues of origin, uncovering marked differences in the spectrum of miRNA targets between cancers and normal tissues.

When compared to standard immunoprecipitation-based approaches, HEAP offers several advantages. First, the covalent nature of the interaction between the HaloTag and the HaloTag ligands simplifies the isolation of Ago2-miRNA-RNA complexes and removes the intrinsic variability of immunoprecipitation-based approaches. This feature is illustrated by the highly reproducible identification of miRNA-binding sites in murine embryonic stem cells, in developing embryos, in murine tissues and in tumors. Second, the conditional Cre-loxP-based nature of the Halo-Ago2 mouse strain enables the purification of Ago2-containing complexes and the identification of miRNA-RNA interaction sites from a specific subset of cells, thus bypassing the need for microdissection and cell purification using cell surface markers. As proof of concept, we demonstrate this ability by mapping miRNA-RNA interactions in three mouse models of human cancers driven by distinct combina-

tions of oncogenes and tumor suppressor genes. We predict that the systematic application of HEAP will allow the construction of a detailed map of miRNA targets across tissues and cell types in mice.

Despite these advantages, some limitations of the HEAP method should be considered when planning experiments. First, as is true for any tagged protein, the presence of the HaloTag may have functional consequences. The reduced viability of the Halo-Ago2 homozygous animal we have observed does indicate that Halo-Ago2 is not entirely functionally identical to Ago2, perhaps due to reduced stability or to a subtle impairment of miRISC assembly and activity. Thus, it will be important to experimentally evaluate the functional relevance of individual miRNA-RNA interactions identified using this approach. Second, although the conditional nature of the Halo-Ago2 allele is ideally suited for the direct identification of miRNA targets in rare cell populations within a tissue, the HEAP method requires a relatively large number of cells (ideally 1×10^7 cells or more) to produce robust results, and in some cases, it may be therefore necessary to pool tissues from multiple animals.

In conclusion, the HEAP method and the Cre-inducible Halo-Ago2 mouse strain described here, combined with the growing array of strains expressing Cre in a temporally and spatially restricted fashion, will facilitate the generation of detailed maps of miRNA-RNA interactions *in vivo* under physiological and pathological conditions.

Future directions

The tool and datasets presented in this thesis provide great opportunities to carefully study miRNA regulations in a variety of physiologic contexts. The initial characterization of these datasets also raises lots of interesting questions and

opens up new possibilities.

We have identified a large number of miRNA binding sites outside of 3'UTRs. The biological functions of these binding sites remain to be investigated. One of such examples is a miR-17/20-5p binding site found in the 5'UTR of *Tnrc6b*, one of the genes encoding GW182. This site falls in the 5'UTR of a short isoform of *Tnrc6b* expressed from an alternative promoter and was observed across several cell/tissue types examined in this study, including mESCs, embryos, gliomas and lung adenocarcinomas (**APPENDIX IV**). In particular, this site is adjacent to the start codon of the mRNA and potentially affects translation initiation. Preliminary analysis has suggested that, at least in mESCs, 5'UTR binding sites are not associated with consistent target mRNA expression changes (data not shown). This observation raises a series of interesting questions. Whether the two isoforms of *Tnrc6b* are functionally different? Why does miRISC target mRNA encoding the miRISC resident protein? Does this regulation work as a universal negative feedback mechanism in the miRNA pathway or, on the contrary, facilitate the assembly of miRISC by bringing translating *Tnrc6b* close to the miRNA pathway machineries? Further efforts are required to carefully examine the molecular mechanism through which this 5'UTR target site regulates *Tnrc6b* expression or even function.

Although the HEAP method has drawn miRNA target maps in various cell- and tissue-types, it is still computationally challenging to determine the physiologic roles of the binding sites. A possible approach to address this question is to perform functional genetic screens on these sites. CRISPR-assisted large-scale screens have been applied to coding sequences (Shalem et al., 2014; Wang et al., 2014), gene regulatory elements (Korkmaz et al., 2016), but not to miRNA binding sites in 3'UTRs. We envision to design guide RNA libraries to disrupt individual miRNA binding sites in the HEAP libraries and measure cellular responses by

high-throughput sequencing. Mutating a single miRNA binding site in cells offers several advantages as it abolishes the binding of all redundant miRNAs working at the same site and avoids affecting expression of other target genes regulated by the same miRNAs. This approach can be used to determine important miRNA-RNA interaction events for mESC maintenance and tumor cell survival *in vitro*. When applied to *in vivo* tumor models, it is useful for dissecting miRNA regulations during tumor initiation and progression.

The HEAP workflow presented in this study is only a prototype. The efficiency of this method can be further improved by optimizing individual reactions and purification steps. In addition, we emphasize that the HEAP protocol can be easily modified to accommodate the many variations of the basic HITS-CLIP strategy, including those using ligation to generate chimeric reads between the mature miRNA and its target (CLASH, CLEAR-CLIP), and those designed to identify the crosslinking site at single base resolution (PAR-CLIP, iCLIP, eCLIP).

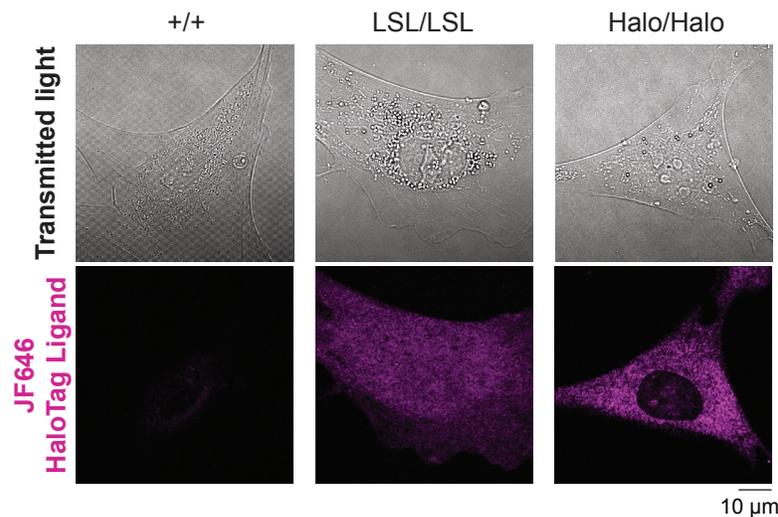


Figure 5.1: Live cell imaging of the Halo-Ago2 fusion protein
Confocal imaging of MEFs of indicated genotypes. MEFs were incubated with 200 nM Janelia Fluor 646 HaloTag ligand 1hr prior to imaging. +/+ : $Ago2^{+/+}$; LSL/LSL: $Ago2^{Halo-LSL/Halo-LSL}$; Halo/Halo: $Ago2^{Halo/Halo}$.

Although in this study we have focused exclusively on the identification of miRNA-RNA interactions in cells and tissues, the conditional Halo-Ago2 mouse strain we have developed could prove useful for the biochemical characterization of Ago2-containing protein complexes *in vivo* and for imaging studies (**Figure 5.1**). Notably, fluorescent HaloTag ligands have been successfully used recently for super-resolution imaging of Halo-tagged proteins (Grimm et al., 2015). When applied to cells and tissues expressing the Halo-Ago2 knock-in allele, this strategy could provide insights into the subcellular localization and dynamics of this important RBP under different conditions and in response to external and internal cues.

In summary, the tools and the datasets we have generated will serve as a valuable resource to the scientific community and will facilitate the characterization of miRNA functions under physiological and pathological conditions.

APPENDIX I: Step wise protocol to generate HEAP and input control libraries

HEAP Library preparation

Step 1. Cell preparation and UV crosslinking

1. Harvest and irradiate cells.

For mESCs

- a) Dissociate mESCs from culture dishes and wash with cold PBS.
- b) Resuspend cells in cold PBS in a 10 cm² dish and place on ice.
- c) Crosslink cells by placing the dish on ice in a Spectroline UV crosslinker. Irradiate cells at dose 400 mJ/cm²

For fresh tissues or tumors

- a) Euthanize mice following the standard protocol.
 - b) Resect tissues of interest from mice and place in cold PBS on ice.
 - c) Homogenize tissues with a scalpel.
 - d) Crosslink tissues on ice at dose 400 mJ/cm² for three times.
2. Collect cells or tissues by centrifugation at 4 °C, 1,300 ×g for 10 min.
 3. Remove supernatant. Snap-freeze the pellet on dry ice and store at -80 °C until use.

Step 2. Lysates preparation and RNase digestion.

1. Resuspend pellets in 3 volumes of Mammalian Lysis Buffer containing protease inhibitor cocktail (50×, Promega G6521).
2. Pipette/vortex to mix. Incubate on ice for 15 min.
3. Add 25 µL RQ1 DNase (Promega M6101) per 300 µL lysate. Incubate in a thermomixer at 37 °C, 1,000 rpm for 5 min.
4. Per 300 µL lysates, add 2.5 µL RQ1 DNase and 10 µL RNase A (Affymetrix, 1:50,000 diluted in TBS). Incubate in a thermomixer at 37 °C, 1,000 rpm for 5 min.
5. Pass lysates through a 26-gauge needle to reduce viscosity.
6. Clear lysates by centrifugation at 4 °C, 14,000 rpm for 10 min. Transfer lysates to a new tube and place on ice.

Step 3. Halolink Resin equilibration

1. Homogenize the Halolink resin (Promega G1914) slurry by inversion and dispense into a 15 mL conical tube (300 µL slurry per sample).
2. Wash resin with 4 volumes of Wash/Eq buffer.
3. Collect resin by centrifugation at 800 ×g for 2 min.
4. Repeat step 2-3 twice.
5. Leave the resin in Wash/Eq buffer in a 1.5 mL eppendorf tube until lysates are ready.

Step 4. Halo-Ago2 pulldown and washing

1. Save ~2% lysate for **input control library** preparation. Store at -80 °C until use.
2. Dilute rest of the lysate with TBS at a 3:7 dilution ratio (700 µL TBS added into 300 µL lysates).
3. Remove Wash/Eq buffer from the equilibrated resin and add diluted lysate.
4. Incubate lysate with resin on a tube rotator at room temperature for a total of 1.5 hr. If the volume of diluted lysate is greater than 1 mL, collect the resin by centrifugation (at 800 ×g for 2 min) and reload the resin with the same sample.
5. Wash resin with 100 µL SDS elution buffer. Rotate at room temperature for 30 min. Collect resin by centrifugation at 800 ×g for 2 min.
6. Wash resin 3× with 1 mL LiCl wash buffer, 2× with 1 mL PXL (1×) buffer, 2× with 1 mL PXL (5×) buffer and 2× with 1 mL PNK buffer. Collect resin by centrifugation in between.

Step 5. Dephosphorylation

1. Prepare the following reaction mix:

Ref	Components	Volume (µL)
Promega M183A	10× Alkaline phosphatase buffer	8
Promega M182A	Calf Intestinal Alkaline Phosphatase	3
Promega N251B	rRNasin	2
	Water	67

2. Remove the residual PNK buffer from the resin and add 80 µL phosphatase reaction mix. Incubate in a thermomixer at 37 °C for 20 min. Mix at 1,000 rpm for 15 s every 2 min.
3. Wash resin 2× with 1 mL PNK-EGTA buffer and 2× with 1 mL PNK buffer.

Step 6. 3' RNA linker ligation

1. Prepare the following reaction mix:

Ref	Components	Volume (µL)
NEB B0216S	10× T4 RNA ligase buffer	8
	BSA (0.2 µg/µL)	8
NEB P0756S	ATP (10 mM)	8
NEB M0204S	T4 RNA ligase 1	3
Promega N251B	rRNasin	2
	RL3 (20 µM)	5
	Water	46

2. Add 80 µL mixture to the resin. Incubate in a thermomixer at 16 °C overnight. Mix at 1,000 rpm for 15 s every 2 min.

Step 7. PNK treatment

1. Next day, wash resin 1× with 1 mL PXL (1×) buffer, 1× with 1 mL PXL (5×) buffer and 3× with 1 mL PNK buffer.
2. Prepare the following reaction mix:

Ref	Components	Volume (μL)
NEB B0201S	10× PNK buffer	8
NEB P0756S	ATP (10 mM)	1
NEB M0201L	T4 PNK	4
Promega N251B	rRNasin	2
	Water	65

3. Add 80 μL mixture to the resin. Incubate in a thermomixer at 37 °C for 20 min. Mix at 1,000 rpm for 15 s every 2 min.
4. Wash resin 3× with 1 mL PNK buffer and 1× with 1 mL Wash/Eq buffer.

Step 8. RNA isolation from resin

1. Dissolve proteinase K (Roche 3115836001) in PK buffer to a final concentration of 4 mg/mL. Pre-warm this solution in a thermomixer at 37 °C, 1,000 rpm for 20 min.
2. Wash the resin with 1 mL PK buffer.
3. Add 200 μL pre-warmed proteinase K solution. Incubate in a thermomixer at 37 °C, 1,000 rpm for 20 min.
4. Add 200 μL PK-Urea solution. Incubate in a thermomixer at 37 °C, 1,000 rpm for 20 min.
5. Add 400 μL Phenol (pH4.3, Sigma P4682) and 130 μL Chloroform (Sigma-Aldrich: 25668). Incubate in a thermomixer at 37 °C, 1,000 rpm for 20 min.
6. Spin at 14,000 rpm at room temperature for 5 min.
7. Transfer the aqueous phase to a siliconized tube.
8. Add 50 μL 3 M Sodium Acetate (pH 5.5, Ambion AM9740) and 0.75 μL Glycogen (5 mg/mL, Ambion AM9510). Mix thoroughly.
9. Add 1 mL cold Ethanol/Isopropanol (1:1 by volume). Precipitate overnight at minus 20 °C.

Step 9. RNA precipitation and 5' RNA linker ligation

1. Precipitate the RNA by centrifugation at 4 °C, 14,000 rpm for 20 min.
2. Wash the pellet once with 1 mL cold 70% Ethanol. Spin at 4 °C, 14,000 rpm for 10 min.
3. Remove the supernatant and air-dry the pellet.
4. Resuspend the RNA pellet in 5.9 μL RT-PCR grade water.
5. Prepare the following reaction mix:

Ref	Components	Volume (μL)
NEB B0216S	10× T4 RNA ligase buffer	1
	BSA (0.2 μg/μL)	1
NEB P0756S	ATP (10 mM)	1
NEB M0204S	T4 RNA ligase 1	0.1
	RL5D-6N (20 μM)	1

6. Add 4.1 μL mixture to the dissolved RNA. Mix. Incubate at 16 °C for 5 hrs.

Step 10. DNase treatment

1. Prepare the following reaction mix:

Ref	Components	Volume (μL)
Promega M198A	RQ1 DNase 10 \times Reaction Buffer	11
Promega N251B	rRNasin	5
Promega M610A	RQ1 DNase	5
	Water	79

- Add 100 μL mixture to each sample. Incubate in a thermomixer at 37 $^{\circ}\text{C}$, 1,000 rpm for 20 min.
- Add 300 μL H_2O , 300 μL Phenol (pH4.3) and 100 μL Chloroform. Vortex to mix.
- Spin at 14,000 rpm at room temperature for 5 min.
- Transfer the aqueous phase to a siliconized tube.
- Add 50 μL 3 M Sodium Acetate (pH 5.5) and 0.75 μL Glycogen (5 mg/mL). Mix thoroughly.
- Add 1 mL cold Ethanol/Isopropanol (1:1 by volume). Precipitate overnight at minus 20 $^{\circ}\text{C}$.

Step 11. Reverse-transcription

- Precipitate the RNA as described in Step 9 (1-3). Resuspend RNA in 10 μL RT-PCR grade water.
- Prepare the following mix. (RT-: control reaction without reverse transcriptase)

Ref	Components	RT+	RT-
	DP3 (10 μM)	1	1
Invitrogen 18427-013	10 mM dNTP mix	1	1
	Water	3	9

- Add 5 μL of RT+ mixture to 8 μL of RNA and 11 μL of RT- mixture to 2 μL of RNA.
- Incubate at 65 $^{\circ}\text{C}$ for 5 min and 4 $^{\circ}\text{C}$ for 1 min.
- Prepare the following reaction mix:

Ref	Components	RT+	RT-
Invitrogen 18080044	5X First-Strand buffer	4	4
Invitrogen 18080044	0.1M DTT	1	1
Invitrogen 10777019	RNaseOUT (40U/ μl)	1	1
Invitrogen 18080044	SuperScript III Reverse transcriptase (200U/ μl)	1	0
	Water	0	1

- Add 7 μL enzyme-buffer mixture to each corresponding sample.
- Incubate the reaction mixtures in a thermocycler using program:

50 $^{\circ}\text{C}$	45 min
55 $^{\circ}\text{C}$	15 min
90 $^{\circ}\text{C}$	5 min
4 $^{\circ}\text{C}$	∞

Step 12. First PCR to determine the optimal amplification cycle

OPTION 1: diagnostic PCR

1. Prepare the following PCR reaction mix:

Ref	Components	Volume (µL)
Invitrogen 12344040	Accuprime Pfx SuperMix	13.5
	DP5 (20 µM)	0.375
	DP3 (20 µM)	0.375
	cDNA*	1

* Prepare separate mixtures for control samples (RT- and water).

2. Amplify the cDNAs on a thermocycler using the following program with several different cycles. For example, one can start with 18, 22 and 26 cycles.

Temperature	Time	
95 °C	2 min	
95 °C	20 s	X cycles
58 °C	30 s	
68 °C	20 s	
68 °C	5 min	
4 °C	∞	

3. Mix the PCR product with 2x TBE-Urea sample buffer (Invitrogen LC6876).
4. Load the same sample amplified with different number of cycles next to each other. Run PCR products on a 15 % TBE-Urea polyacrylamide gel (Invitrogen EC6885), along with the 25-bp DNA step ladder (Promega G4511), following standard protocol.
5. Stain the gel in 1x SYBR gold nucleic acid gel stain (Invitrogen S11494) in 1 X TBE for 10 min.
6. Visualize the PCR products under UV. Expected size for miRNAs is ~65 bp and expected size range for miRNA targets is 75~200 bp.
7. Determine the optimal amplification cycle.

OPTION 2: real-time PCR

1. Prepare the following reaction mix and load into a 384-well PCR plate.

Ref	Components	Volume (µL)
Invitrogen S7563	50X SYBR Green	0.1
Invitrogen 12344040	Accuprime Pfx SuperMix	9.1
	DP5 (20 µM)	0.25
	DP3 (20 µM)	0.25
	cDNA	0.3

* Prepare triplicates for each sample, 1 µL cDNA is added to a master mix of 30 µL. Therefore, the optimal cycle number should be N-1. N is the optimal cycle determined by real-time PCR.

- Monitor the amplification under a real-time thermocycler using the following program:

Temperature	Time	
95 °C	2 min	
95 °C	20 s	30 cycles
58 °C	30 s	
68 °C	20 s	

- Determine the highest amplification cycle N before the SYBR green signal reaches a plateau. The optimal amplification cycle is N-1.

Step 13. Library preparation – pre-amplification of cDNA libraries

- Amplify the cDNA library with the best cycle determined from previous diagnostic/real-time PCR. Prepare 6 PCR reactions for each library to be made using recipe described in Step 12, OPTION1-1.
- Load samples into 15% TBE-Urea gels. Resolve the miRNA and target bands on gel. Stain the gel with SYBR gold nucleic acid gel stain.
- Excise the miRNA band (~65 bp) and target (75~200 bp, usually a smear) separately from the gel.
- Cut the bands into slices and place the gel pieces in a 0.5 mL eppendorf tube with a hole on the bottom.
- Place the 0.5 mL tube in a 2 mL eppendorf tube. Pass the gel through the hole by centrifugation at 13,000 rpm, 4 °C for 1 min.
- Weigh the gels and add 1-2 volumes of diffusion buffer.
- Incubate the gel pieces with diffusion buffer in a thermomixer at 55 °C, 1,000 rpm for 30 min.
- Centrifuge at 14,000 rpm, 4 °C for 1 min to clear the diffusion buffer.
- Pass the supernatant through a Nanosep column (0.2 µm, PALL corporation ODM02C34).
- Determined the volume of supernatant and add 3 volume of buffer QG (Qiagen MinElute Gel Extraction Kit 28606).
- Pass the samples through the Qiagen MinElute spin columns. Wash two times with buffer PE.
- Elute DNA with 10~20 µL H₂O

Step 14. Library construction – Introducing sequencing adaptors

- Design library multiplexing strategy and assign different barcodes to samples to be run on the same lane in an illumina flow-cell.
- Prepare the following PCR reaction mixtures:

Ref	Components	Volume (µL)
Invitrogen 12344040	Accuprime Pfx SuperMix	27
	DSFP5 (20 µM)	0.5
	DP3-Barcode (20 µM)	0.5
	Eluted DNA	3

*Prepare three reactions for each library.

- Amplify each library with different number of cycles using the following program.

Temperature	Time	
95 °C	2 min	
95 °C	20 s	X cycles (x = 5, 7, 9...)
58 °C	30 s	
68 °C	40 s	
68 °C	5 min	
4 °C	∞	

- Load the same library amplified with different cycles side-by-side onto a 6% TBE polyacrylamide gel (Invitrogen EC6265) and run the gel following standard protocol.
- Select the best amplification cycle (usually the lowest cycle) and cut the PCR product from the gel.
- Cut the gel into pieces and add 300 µL water to elute DNA. Incubate the gel with water on a rotator at 4 °C overnight.
- Next day, remove gel pieces from water by passing it through a Nanosep column.
- Precipitate DNA by adding 30 µL 3 M Sodium Acetate (pH5.5), 2 µL glycogen and 2 µL 0.1× NF-Pellet Paint (Novagen 70748-3) and 975 µL absolute Ethanol.
- Pellet DNA by centrifugation at 4 °C, 14,000 rpm for 20 min.
- Wash pellet with 500 µL 70% Ethanol.
- Air-dry the pellet and resuspend in 15 µL water.
- Submit the DNA libraries to the Integrated Genomics Operation Core at Memorial Sloan Kettering Cancer Center for quality-control, quantification, library pooling and high-throughput sequencing.

Input Control Library

Step 1. Dephosphorylation of RNA 3'ends

- Prepare the following reaction mix:

Ref	Components	Volume (µL)
Promega M183A	10× Alkaline phosphatase buffer	2.5
Promega M182A	Alkaline phosphatase, Calf Intestine	2.5
Promega N251B	rRNasin	0.5
	Water	9.5

- Add 15 µL mixture to 10 µL lysates saved in Step 4 before Halo-Ago2 pulldown.
- Incubate in a thermomixer at 37°C, 1,000 rpm for 20min.

Step 2. T4 PNK treatment

- Prepare the following reaction mix:

Ref	Components	Volume (μL)
NEB B0201S	10× PNK buffer	10
NEB P0756S	ATP (10mM)	1.25
NEB M0201L	T4 PNK	5
Promega N251B	rRNasin	2.5
	Water	56.25

- Add 75 μL mixture directly to each sample (25 μL).
- Incubate in a thermomixer at 37°C, 1,000 rpm for 20min.

Step 3. RNA clean-up

Clean up input RNA using MyOne Silane Beads (Thermo 37002D) (Adapted from eCLIP protocol).

a) Prepare beads:

- Dispense 20 μL MyONE Silane beads per sample into eppendorf tubes, magnetically separate and remove supernatant.
- Wash 1x with 900 μL RLT buffer (Qiagen 79216).
- Resuspend beads in 300 μL RLT buffer per sample.

b) Bind RNA:

- Add beads in 300 μL RLT buffer to each sample. Mix.
- Add 10 μL 5M NaCl and 615 μL Absolute Ethanol.
- Rotate at room temperature for 15 min.

c) Wash beads:

- Wash beads with 1 mL 75% Ethanol, pipette resuspend and move the suspension to a new tube.
- After 30 s, magnetically separate and remove supernatant.
- Wash beads 2× with 1 mL 75% Ethanol.
- Spin the tube, magnetically separate and remove supernatant.
- Air-dry the beads for 5 min.

d) Elute RNA:

- Resuspend beads in 10 μL H₂O.
- Magnetically separate.
- Transfer H₂O to a new tube.

Step 4. 3' RNA linker ligation

- Prepare the following reaction mix:

Ref	Components	Volume (μL)
NEB B0216S	10× T4 RNA ligase buffer	2
	BSA (0.2 μg/μl)	2
NEB P0756S	ATP (10 mM)	2
NEB M0204S	T4 RNA ligase 1	0.75
Promega N251B	rRNasin	0.5
	RL3 (20 μM)	1.25
	Water	1.5

2. Add 10 μ L mixture to 10 μ L eluted RNAs.
3. Incubate in a thermomixer at 16°C overnight.

Step 5. RNA clean-up

Next day, clean up RNA using MyONE silane beads.

a) Prepare beads:

1. Dispense 20 μ L MyONE Silane beads per sample into eppendorf tubes, magnetically separate and remove supernatant.
2. Wash 1x with 900 μ L RLT buffer.
3. Resuspend beads in 61.6 μ L RLT buffer per sample.

b) Bind RNA:

1. Add beads in 61.6 μ L RLT buffer to each sample. Mix.
2. Add 61.6 μ L Absolute Ethanol.
3. Incubate at room temperature for 15 min. Pipette mix every 3~5 min.

c) Wash beads:

1. Wash beads with 1 mL 75% Ethanol, pipette resuspend and move the suspension to a new tube.
2. After 30 s, magnetically separate and remove supernatant.
3. Wash beads 2 \times with 1mL 75% Ethanol.
4. Spin the tube, magnetically separate and remove supernatant.
5. Air-dry the beads for 5 min.

d) Elute RNA:

1. Resuspend beads in \sim 6 μ L H₂O, let it sit for 5 min.
2. Magnetically separate.
3. Transfer 5.9 μ L supernatant to a new tube.

Step 6. 5' RNA adaptor ligation

(See Step 9-5 in HEAP library preparation)

Step 7. DNase treatment

(See Step 10 in HEAP library preparation)

Step 8. Reverse transcription for input RNA

(See Step 11 in HEAP library preparation)

Step 9. First PCR to determine the optimal amplification cycle

Determine the optimal amplification conditions for input control libraries as instructed in Step 12 in HEAP library preparation. Based on experience, the optimal cycles are often between 13 to 18 cycles.

Step 10. Library preparation – pre-amplification of input cDNA libraries

Follow procedures in Step 13 in HEAP library preparation. However, a minor change is applied to the gel purification step. To prepare “size-matched” input control libraries for the corresponding HEAP libraries, extract PCR products between 75 and 200 bp from the 15% TBE-Urea gel.

Step 11. Library construction – Introducing sequencing adaptors
(See Step 14 in HEAP library preparation)

Buffer Recipes

1× PBS

137 mM	NaCl
2.7 mM	KCl
10 mM	Na ₂ HPO ₄
1.8 mM	KH ₂ PO ₄

5× PBS

685 mM	NaCl
13.5mM	KCl
50 mM	Na ₂ HPO ₄
9 mM	KH ₂ PO ₄

1× TBS

100 mM	Tris-HCl (pH7.5)
150 mM	NaCl

Mammalian Lysis Buffer (Promega)

50 mM	Tris-HCl (pH7.5)
150 mM	NaCl
1%	Triton X-100
0.1%	Na deoxycholate

Wash/Eq

0.05% IGEPAL CA-630 in 1xTBS

SDS Elution Buffer (10 mL)

0.1%	SDS
50 mM	Tris-HCl (pH7.5)

LiCl Wash Buffer

100 mM	Tris-HCl (pH8.0)
500 mM	LiCl
1%	IGEPAL CA-630
1%	Na deoxycholate

PXL (1×)

In 1× PBS, add:

0.1%	SDS
0.5%	Na deoxycholate
0.5%	IGEPAL CA-630

PXL (5×)

In 5× PBS, add:

0.1%	SDS
0.5%	Na deoxycholate
0.5%	IGEPAL CA-630

1× PNK Buffer

50 mM	Tris-HCl (pH7.4)
10 mM	MgCl ₂
0.5%	IGEPAL CA-630

1× PNK + EGTA

50 mM	Tris-HCl (pH7.4)
20 mM	EGTA
0.5%	IGEPAL CA-630

PK Buffer (Proteinase K)

100 mM	Tris-HCl (pH7.5)
50 mM	NaCl
10 mM	EDTA (pH8.0)

1× PK Buffer/7M Urea (prepare FRESH each time)

100 mM	Tris-HCl (pH7.5)
50 mM	NaCl
10 mM	EDTA (pH8.0)
7M	Urea

Diffusion buffer

0.5 M	Ammonium acetate
10 mM	Magnesium acetate
1 mM	EDTA (pH8.0)
0.1%	SDS

APPENDIX II: Computational pipeline used to process HEAP and input control libraries

Step 1. Barcode removal

Script:

```
#!/usr/bin/env python2
import multiprocessing
import functools
from Bio import Seq, SeqIO
import gzip
import csv
import os
import glob

def attach_barcode( file_name, barcode_len ):
    input_file = "%s.fastq.gz" % file_name
    input_handle = gzip.open( input_file, "rb" )
    output_file = "%s_barcode.fastq.gz" % file_name
    output_handle = gzip.open( output_file, "wb" )

    for record in SeqIO.parse( input_handle, "fastq" ):
        record.description = ""
        sequence = str( record.seq )
        tag = sequence[ :barcode_len ]
        record.id = record.id + "." + tag
        record = record[ barcode_len: ]
        SeqIO.write( record, output_handle, "fastq" )

    print "%s Complete!\n" % file_name
    input_handle.close()

file_names = glob.glob("**fastq.gz")
file_names = [ x[:-9] for x in file_names ]

max_num_process = len( file_names )
pool = multiprocessing.Pool( max_num_process )
test = pool.map( functools.partial( attach_barcode, barcode_len = 7 ), file_names )
```

Usage: Barcode length (“barcode_len”) is 7. The 6-nt degenerate barcode and the “G” introduced by RL5D-6N is to be removed.

Step 2. Adaptor removal and quality trimming

Script:

```
cutadapt -a GTGTCAGTCACTTCCAGCGGGATCGGAAGAGCACACGTCTGAACTC-  
CAGTCAC -m 18 -q 20 -o filename.trimmed.fastq.gz filename.fastq.gz
```

Step 3. Alignment (for HEAP mRNA and input control libraries only)

Script:

```
STAR \  
  --genomeLoad NoSharedMemory \  
  --genomeDir mm10_star/ \  
  --readFilesIn filename.trimmed.fastq.gz \  
  --runThreadN 2 \  
  --alignIntronMin 70 \  
  --alignIntronMax 100000 \  
  --outSAMtype BAM SortedByCoordinate \  
  --outFilterMultimapNmax 1 \  
  --outFilterMultimapScoreRange 0 \  
  --outFilterMismatchNmax 5 \  
  --outFileNamePrefix filename/ \  
  --outStd Log \  
  --readFilesCommand zcat \  
  --outReadsUnmapped Fastx
```

Usage: “mm10_star” genomic index was preliminarily generated using command “STAR --runMode genomeGenerate”

Step 4. PCR duplicate removal

Script:

```
#!/usr/bin/env perl
```

```
# Collapse CLIP reads mapped to the same position & with the same barcode
use warnings;
use 5.014;
```

```
# Calculate the range that a read span
sub realen {
    my $realen = 0;
    while( $_[0] =~ m/(\d+)([MIDNS])/g ){
        my $val = $1;
        my $code = $2;
        given($code) {
            when ('M') { $realen = $realen + $val;}
            when ('I') { $realen = $realen + 0;}
            when ('D') { $realen = $realen + $val;}
            when ('N') { $realen = $realen + $val;}
            when ('S') { $realen = $realen + 0;}
        }
    }
    $realen;
}
}
```

```
$_ = "\n";
# open FILE, $ARGV[0] or print "File $ARGV[0] doesn't exist.";
# Fields I need
my %reads = ();
my $tag = "";
my $chr = "";
my $strand = "";
my $pos = 0;
my $nh = "";
my %flag = ();
my $key = "";
while(<>){
    chomp;
    # Do nothing with the header lines
    if ( substr($_,0,1) eq "@"){
        print;next;
    }
    # Now this line is a real read
    my @fields = split "\t";
    # Skip unmapped reads
    next if($fields[2] eq '*');
    # Fields to keep:
    $tag = (split( ':', $fields[0] ))[-1];
    $chr = $fields[2];
    $strand = ( ($fields[1] & 0x10) == 0 )?"+":"-";
    # Position that the first nt of a read is mapped to
```

```

# Positive strand: original pos
# Reverse: go forward a bit
$pos = ( ($fields[1] & 0x10) == 0 )?($fields[3]):( $fields[3] + realen($fields[5]) - 1 );
$flag{ ( split ":", $fields[$_] ) [0] } = ( split ":", $fields[$_] ) [-1] foreach ( 11 .. $#fields );
$nh = $flag{NH};
# die "Parsing the wrong field!\n" if ( ( split ":", $fields[13] ) [0] ne "X0" );
$key = join ":", $chr, $strand, $pos, $tag;
# $pos = ( ($fields[1] & 0x10) == 0 )?($fields[3]):( $fields[3] + realen($fields[5]) - 1);
# Multi-hit reads have lower priority
$reads{$key} = join "\t", @fields if( $nh == 1 or !exists( $reads{$key} ) )
}
foreach( keys %reads ){
    print $reads{$_};
}
# close FILE;

```

Step 5. HEAP miRNA library alignment

Script:

```

bowtie2 --no-unal -p 7 -x mouse_hairpin/mouse_hairpin -U filename.trimmed.fastq.gz |
samtools view -bS - | samtools sort -o filename.bam

```

Usage: “mouse_hairpin” miRNA genome index was built from 1,915 murine pre-miRNA sequences from miRbase version 21.

Step 6. miRNA abundance measurement

Script:

```

require( GenomicAlignments )
require( stringr )
require( parallel )

mir <- readRDS( “hairpin_info.rds” )
mir.end <- resize( resize( mir, 1, fix = ‘end’ ) , 9, fix = ‘center’ )
mir.end <- split( mir.end, mir.end$mature )

fl <- Sys.glob( “*.bam” )
aln <- lapply( fl, readGAlignments, param = ScanBamParam( tag = “NM” ) )

```

```

mir.counts <- mclapply( aln, function( aln, mir.end ){
  aln <- aln[ mcols( aln )$NM <= 1]
  gr <- resize( granges( aln ), 1, fix = "end" )

  counts <- countOverlaps( mir.end, gr )
  return( counts )
} , mir.end, mc.cores = 7 )
mir.counts <- do.call( 'cbind', mir.counts )
colnames( mir.counts ) <- substr( basename(fl), 1, nchar(basename(fl)) - 4 )
write.csv( mir.counts, file = "mature_mir_counts_all.csv" )

```

Usage:

1. Call this Rscript from command line.
2. "hairpin_info.rds" is a GenomicRanges object containing coordinates of mature miRNAs relative to their pre-miRNAs.

Step 7. Peak calling

RScript:

```

peak.data <-
  findPeaks(bamfiles,
            names(filenamees),
            condition = c(1, 0, ...),
annot.order = c("utr3", "utr3*", "utr5", "utr5*",
                "exon", "intron"),
            exclude.mirna.peaks = TRUE,
            genome = "mm10",
            bandwidth = bandwidth,
            count.threshold = count.threshold,
            extend.slice = extend.slice,
            count.exons.only = FALSE,
            extend.peaks.in.genes = extend.peaks.in.genes,
            transform.chr.names = TRUE,
nthreads = 14)

```

Usage:

1. "bamfiles" is a vector containing all BAM files, including HEAP mRNA library and

input library.

2. “filenames” is a vector of sample names matched to BAM files.
3. The “condition” parameter specifies whether the BAM file is a HEAP mRNA library (1) or an input library (0).
4. “bandwidth”, “extend.slice”, “count.threshold” and “extend.peaks.in.genes” parameters used in each analysis were described in the Methods session.

APPENDIX III: Oligonucleotides

A. Oligonucleotides used in generating HEAP libraries

Name	Sequence
DSFP5	5'-AATGATACGGCGACCACCGACTATGGATACTTAGTCAGGGAGGACGAT-GCGG-3'
SSP1	5'-CTATGGATACTTAGTCAGGGAGGACGATGCGG-3'
DP5	5'-AGGGAGGACGATGCGG-3'
DP3	5'-CCGCTGGAAGTGACTGACAC-3'
RL3	5'-P-GUGUCAGUCACUCCAGCGG 3'-puromycin
RL5D-6N	5'-OH AGGGAGGACGAUGCGNNNNNNG 3'-OH

B. Oligonucleotides used in multiplexing HEAP libraries

Name	TruSeq Index	Sequence
DP3_Barcode 1	13	CAAGCAGAAGACGGCATAACGAGATIGTTGACTGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 2	14	CAAGCAGAAGACGGCATAACGAGATACGGAACTGTGACTG-GAGTTCAGACGTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 3	15	CAAGCAGAAGACGGCATAACGAGATTCTGACATGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 4	16	CAAGCAGAAGACGGCATAACGAGATCGGGACGGGTGACTG-GAGTTCAGACGTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 5	18	CAAGCAGAAGACGGCATAACGAGATGTGCGGACGTGACTG-GAGTTCAGACGTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 6	19	CAAGCAGAAGACGGCATAACGAGATCGTTTCACGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 7	20	CAAGCAGAAGACGGCATAACGAGATAAGGCCACGTGACTG-GAGTTCAGACGTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 8	21	CAAGCAGAAGACGGCATAACGAGATTCGGAAACGTGACTG-GAGTTCAGACGTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 9	22	CAAGCAGAAGACGGCATAACGAGATTACGTACGGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 10	23	CAAGCAGAAGACGGCATAACGAGATATCCACTCGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 11	25	CAAGCAGAAGACGGCATAACGAGATATATCAGTGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC
DP3_Barcode 12	27	CAAGCAGAAGACGGCATAACGAGATAAAGGAATGTGACTGGAGTTCAGAC-GTGTGCTCTTCCGATCCCGCTGGAAGTGACTGACAC

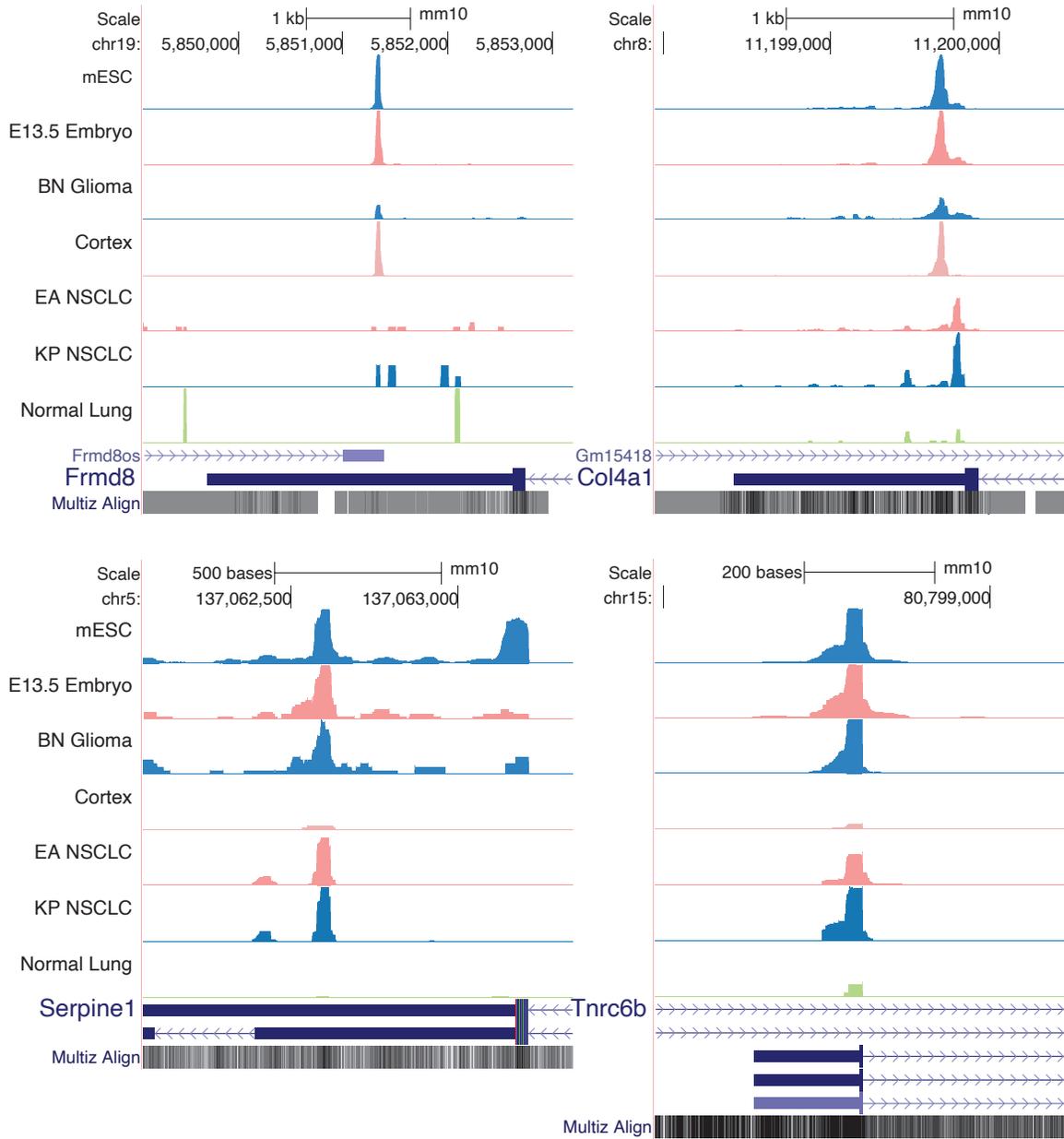
DP3_Barcode 13	1	CAAGCAGAAGACGGCATAACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 14	2	CAAGCAGAAGACGGCATAACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 15	3	CAAGCAGAAGACGGCATAACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 16	4	CAAGCAGAAGACGGCATAACGAGATIGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 17	5	CAAGCAGAAGACGGCATAACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 18	6	CAAGCAGAAGACGGCATAACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 19	7	CAAGCAGAAGACGGCATAACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC
DP3_Barcode 20	8	CAAGCAGAAGACGGCATAACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCCTGGAAGTGACTGACAC

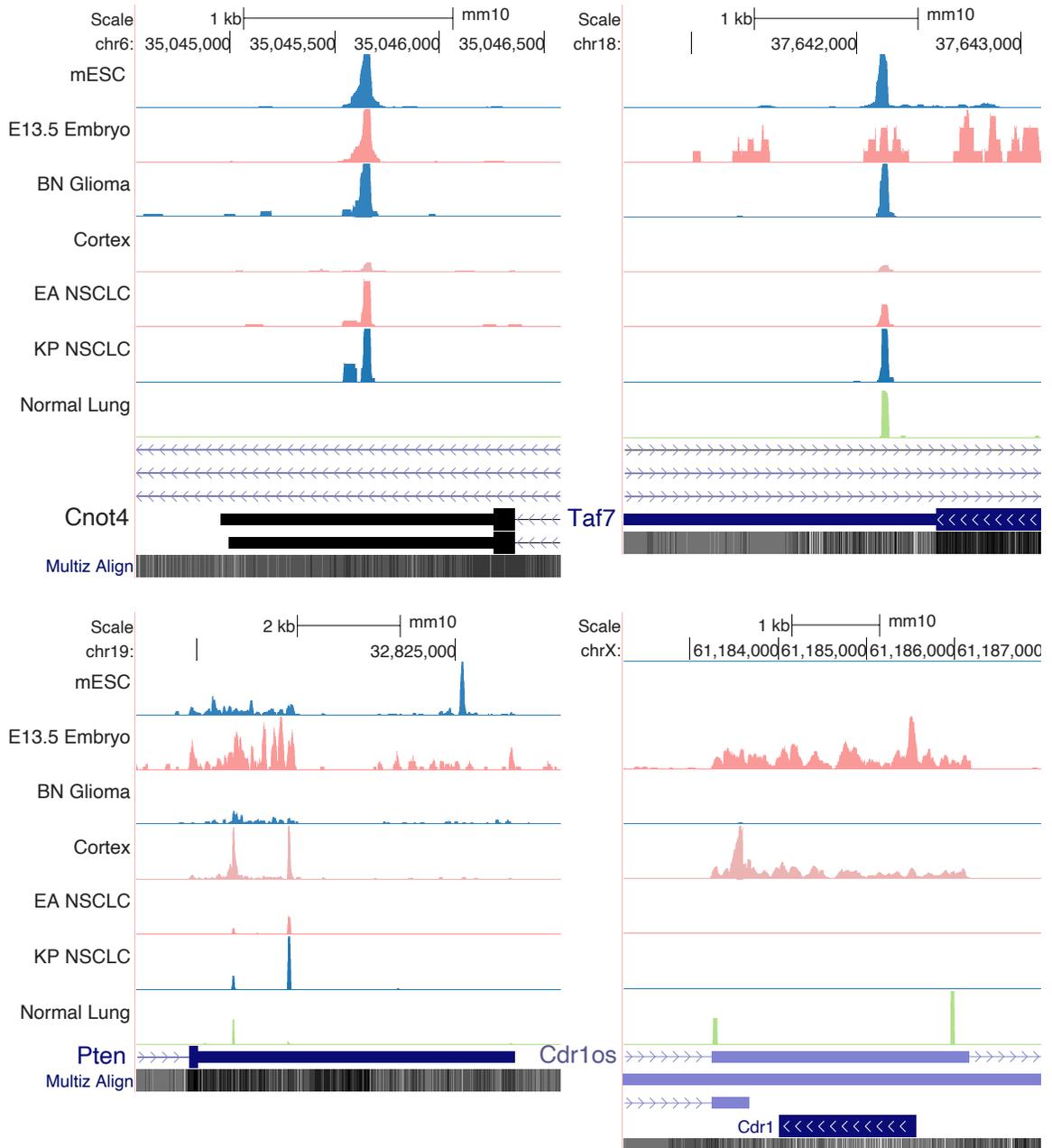
C. Other oligonucleotides

Name	Sequence
Lefty2 3'UTR sgRNA1	TAAGTGCTTAAACTGGGTGA
Lefty2 3'UTR sgRNA2	CCCAGTTTAAGCACTTACAT
Lefty2 donor template 1	AGATGTATTCTCAGTGAGCTTGTCTTAAGTCTCGT-CAGACCTTTGCTCTACAGTCTTGGTTTTCTTGTTCATACCCAGTTAATCGCGTACATGGGTAAATCATGTCACTCCAGTAGGACACACTGACCCCACTTAGC-CAAGGACATGGCTATGCAGTGAACAGG
Lefty2 donor template 2	GAGCTTGTCTTAAGTGTCTCTCGTCAGACCTTTGCTCTACAGTCTTGTTTTCTTGTCCATACCCAGTTAATCGCGTACATGGATAAATCATGTCACTC-CAGTAGGACACACTGACCCCACTTAGCCAAGGACATGGCTATGCAGTGAA-CAGGTTTCGCA
Dicer1 deletion sgRNA1	ATCGACACCACCATGCGGCT
Dicer1 deletion sgRNA2	GGAGAATCAGTCCCGGATTG
Forward primer to amplify Pten 3'UTR	AGCGCGTGCTGAAGAACGAGCAGTAATTCTAGGCGATCGCTTCACGTCCTAC-CCCTTTGC
Reverse primer to amplify Pten 3'UTR	CCAGCGGCCGCTCTAGGTTTAAACGAATTCCTGGGCTCGAGACCTGGC-CCTCTGTGTTAC
Forward primer to amplify Adb2 3'UTR	AGCGCGTGCTGAAGAACGAGCAGTAATTCTAGGCGATCGCAGACCCCTCCTTGACAGGAC
Reverse primer to amplify Adb2 3'UTR	CCAGCGGCCGCTCTAGGTTTAAACGAATTCCTGGGCTCGACACTCATCGGT-CACGACACA

Synthetic Taf7 3'UTR (wild-type)	AGCGCGTGCTGAAGAACGAGCAGTAATTCTAGGCGATCGCTTTGGT-GCTAAGATCAGTAATAAGCTAGCCTACAGGCAGTGAAGTTAATTTCTTTTAAAGAAATATATTTTTTTTTTACCTCAGCAAGTTTCCATGTTACAGAAATGCAAGTTTTTCATAAAGGTTGTTTCTTATACAATAAAAGCTACTAATTTTCACTGTTTATGATGTTGTTCTATGAATACTGAAAGTTGAGGTGGTTCATGGAGTTTAAATGACTGAATGGTGTAGATCAAGATAAATCCTGTTCTTCGAGCCCGGGAATTCGTTAACCTAGAGCGGCCGCTGG
Synthetic Taf7 3'UTR (mutant)	AGCGCGTGCTGAAGAACGAGCAGTAATTCTAGGCGATCGCTTTGGT-GCTAAGATCAGTAAAATGGTAGCCTACAGGCAGTGAAGTTAATTTCTTTTAAAGAAATATATTTTTTTTTTACCTCAGCAAGTTTCCATGTTACAGAAATGCAAGTTTTTCATAAAGGTTGTTTCTTATACAATAAAAGCTACTAATTTTCACTGTTTATGATGTTGTTCTATGAATACTGAAAGTTGAGGTGGTTCATGGAGTTTAAATGACTGAATGGTGTAGATCAAGATAAATCCTGTTCTTCGAGCCCGGGAATTCGTTAACCTAGAGCGGCCGCTGG
Forward primer to mutate Adrb2 let-7 binding site	/5Phos/CTTCGTGACTGGTCAAGTATTAAGAATG
Reverse primer to mutate Adrb2 let-7 binding site	/5Phos/ATACATGAAAACATCATCCAGAC
Forward primer to mutate Pten miR-29 binding site	/5Phos/CGCGGTAGACAAGGCAGCTAGAGTG
Reverse primer to mutate Pten miR-29 binding site	/5Phos/AATATGCTCTTTAAAATGCACAAAC

APPENDIX IV: Example peaks in HEAP libraries





APPENDIX V: Generation of the *Lefty2* mutant mice

Introduction

Validation of the interaction between miR-291-3p and the 3'UTR of *Lefty2* drove us to further interrogate the biological function of this interaction in mouse development, as *Lefty2* is involved in regulating mesoderm development and lefty-right patterning (Meno et al., 1999; Meno et al., 2001). *Lefty2*, along with *Lefty1* and *Nodal*, encodes secreted proteins belonging to the TGF- β superfamily of proteins (Meno et al., 1996; Zhou et al., 1993). In mouse embryos, *Lefty2* is normally expressed in nascent mesoderm around E7.0 and is restricted to the left lateral plate mesoderm (LPM) at early somite stage in response to *Lefty1* and *Nodal* signaling (Meno et al., 1997; Meno et al., 1998; Saijoh et al., 2000). Deletion of *Lefty2* leads to gastrulation defects (Meno et al., 1999). An asymmetric enhancer (ASE) upstream of *Lefty2* confers the asymmetric expression of *Lefty2* in LPM (Saijoh et al., 1999; Saijoh et al., 2000). Embryos lacking ASE fail to express *Lefty2* at LPM. Although the mutant mice survive to term, they display several visceral defects including left isomerism (Meno et al., 2001). Both phenotypes can be explained by the de-regulation of *Nodal* signaling. *Lefty2* is a feedback inhibitor of *Nodal*, presumably by competing for *Nodal* receptors (Meno et al., 1999). Heterozygous deletion of *Nodal* partially rescue the phenotype observed in the *Lefty2* knockout embryos (Meno et al., 1999). *Nodal* also induces the expression of itself in LPM (Saijoh et al., 2000). The absence of *Lefty2* in LPM causes prolonged *Nodal* expression and bilateral expression of *Pitx2*, a homeobox gene downstream of *Nodal*, that drives side-specific transcription programs essential for visceral organ

patterning (Logan et al., 1998; Piedra et al., 1998; Ryan et al., 1998; Yoshioka et al., 1998).

The miRNA binding site identified in the 3'UTR of *Lefty2* is predicted to be targeted by miRNAs from both the miR-291-3p and miR-17/20-5p families, as their seed sequences overlap by 7-nt. miR-291-3p family members are the first miRNAs induced in mouse zygotes and start decreasing at early gastrulation stage (E6.5) (Medeiros et al., 2011). Expression of miRNAs in the miR-17/20-5p family can be readily detected across cell lineages starting from blastocyst stage (Foshay and Gallicano, 2009). Thus, multiple miRNAs possibly cooperate in regulating the expression of *Lefty2*. We hypothesize that the miR-291-3p binding site in the 3'UTR of *Lefty2* may confer post-transcriptional regulation and influence mesoderm development or left-right patterning of mouse embryos. Supporting this hypothesis is that miR-430, a homolog for mouse miR-291 in zebrafish, regulates Nodal antagonist *lefty*. Masking the miR-430 binding site by complementary morpholinos increases *lefty* (*lft2*) expression and impairs Nodal signaling and mesoderm devel-

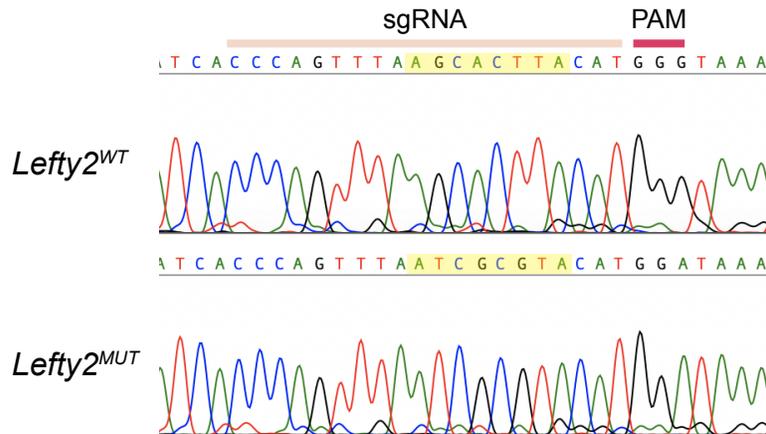


Figure 6.1: DNA sequencing chromatograms of *Lefty2* locus surrounding the miR-291-3p binding site
 sgRNA: single guide RNA; PAM: protospacer adjacent motif.

opment (Choi et al., 2007).

Generation of *Lefty2* miR-291-3p binding site mutant mice

To characterize the function of this miRNA-mRNA interaction event in mouse development, we generated a mouse strain in which the miRNA binding site on *Lefty2* gene is disrupted. To generate the mutation, recombinant Cas9 protein, a guide RNA, and a repair template were delivered to fertilized embryos via microinjection or electroporation, both yielding efficient editing and template integration (**Figure 6.1 and Table 6.1**). Embryo manipulation and mutant strain derivation were performed by the Mouse Genetics Core Facility at Memorial Sloan Kettering Cancer Center.

Approach Editing outcome	Microinjection	Electroporation
Unedited	20	0
NHEJ	20	25
HR	14*	11**
Total	54	36

NHEJ: Non-homologous end joining
HR: Homologous recombination
*, **: Mice harboring small in-dels in the knock-in allele included

Table 6.1: *Lefty2* locus status of mice obtained from CRISPR-mediated gene targeting

We genotyped the mice at weaning age. 14 out of 54 mice generated by microinjection carried the knock-in allele, though small insertions and deletions were introduced around the binding site in several mice (data not shown). All of the mice generated by electroporation carried edited *Lefty2* alleles and 11 of them

harbored the desired mutation (**Table 6.1**). Mice harboring biallelic mutations look normal and undistinguishable from wild-type animals (data not shown), suggesting that, unlike in zebrafish, loss of this miRNA binding site does not cause a strong effect on embryo morphogenesis. Since target expression change caused by a typical miRNA regulation is usually less than 2 fold, de-repression of *Lefty2* may only lead to a subtle change in Nodal signaling. We are currently performing more detailed studies to fully characterize the *Lefty2* mutant strain and carefully examine the expression of *Lefty2* and Nodal pathway components during gastrulation as a result of miR-291-3p binding site loss.

Method

The second guide RNA used in generating *Lefty2*^{MUT} mESCs was selected to generate the *Lefty2* mutant mice. Recombinant Cas9, a synthetic guide RNA and a repair template were delivered to fertilized mouse embryos (C57BL/6J) by microinjection or electroporation. Manipulated embryos were transferred to pseudopregnant females by surgery. Animals were genotyped at weaning stage.

Integration of the repair template creates an additional RsaI site. To screen for animals carrying the knock-in allele, *Lefty2* locus was amplified by PCR (p1, 5'-AGTGAAGGACACCCGAT-3'; p2, 5'-TCCCTGACATGGTACGTTGATG-3'), which generated a 652-bp product. The PCR product was then digested with RsaI, which yields 3 fragments (525 bp, 38 bp and 76 bp) for wild-type allele and 4 fragments (406 bp, 119 bp, 38 bp and 76 bp) for the knock-in allele.

Embryo manipulation and mutant strain derivation were performed by the Mouse Genetics Core Facility at Memorial Sloan Kettering Cancer Center. All studies and procedures were approved by the Memorial Sloan Kettering Cancer Center Institutional Animal Care and Use Committee.

BIBLIOGRAPHY

- Adiliaghdam, F., Basavappa, M., Saunders, T.L., Harjanto, D., Prior, J.T., Cronkite, D.A., Papavasiliou, N., and Jeffrey, K.L. (2020). A Requirement for Argonaute 4 in Mammalian Antiviral Defense. *Cell reports* 30, 1690-1701.e1694.
- Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4.
- Akalin, A., Franke, V., Vlahovicek, K., Mason, C.E., and Schubeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics (Oxford, England)* 31, 1127-1129.
- Auyeung, Vincent C., Ulitsky, I., McGeary, Sean E., and Bartel, David P. (2013). Beyond Secondary Structure: Primary-Sequence Determinants License Pri-miRNA Hairpins for Processing. *Cell* 152, 844-858.
- Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. *Nature* 455, 64-71.
- Baillat, D., and Shiekhattar, R. (2009). Functional Dissection of the Human TNRC6 (GW182-Related) Family of Proteins. *Molecular and Cellular Biology* 29, 4144.
- Bak, M., Silahtaroglu, A., Møller, M., Christensen, M., Rath, M.F., Skryabin, B., Tommerup, N., and Kauppinen, S. (2008). MicroRNA expression in the adult mouse central nervous system. *Rna* 14, 432-444.
- Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281-297.
- Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215-233.
- Bartel, D.P. (2018). Metazoan MicroRNAs. *Cell* 173, 20-51.
- Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science* 336, 233.
- Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes & Development* 20, 1885-1898.

- Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. (2003). Dicer is essential for mouse development. *Nature Genetics* 35, 215-217.
- Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology* 11, R90.
- Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic acids research* 36, D149-D153.
- Boroviak, T., Stirparo, G.G., Dietmann, S., Hernando-Herraez, I., Mohammed, H., Reik, W., Smith, A., Sasaki, E., Nichols, J., and Bertone, P. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development (Cambridge, England)* 145, dev167833.
- Bosson, A.D., Zamudio, J.R., and Sharp, P.A. (2014). Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Mol Cell* 56, 347-359.
- Braun, Joerg E., Huntzinger, E., Fauser, M., and Izaurralde, E. (2011). GW182 Proteins Directly Recruit Cytoplasmic Deadenylation Complexes to miRNA Targets. *Molecular cell* 44, 120-133.
- Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene *hid* in *Drosophila*. *Cell* 113, 25-36.
- Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of MicroRNA-Target Recognition. *PLOS Biology* 3, e85.
- Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., *et al.* (2002). Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences* 99, 15524.
- Chang, T.-C., Wentzel, E.A., Kent, O.A., Ramachandran, K., Mullendore, M., Lee, Kwang H., Feldmann, G., Yamakuchi, M., Ferlito, M., Lowenstein, C.J., *et al.* (2007). Transactivation of miR-34a by p53 Broadly Influences Gene Expression and Promotes Apoptosis. *Molecular cell* 26, 745-752.
- Chekulaeva, M., Mathys, H., Zipprich, J.T., Attig, J., Colic, M., Parker, R., and Filipowicz, W. (2011). miRNA repression involves GW182-mediated recruitment of CCR4-NOT through conserved W-containing motifs. *Nature structural & molecular biology* 18, 1218-1226.
- Cheloufi, S., Dos Santos, C.O., Chong, M.M.W., and Hannon, G.J. (2010). A dicer-inde-

pendent miRNA biogenesis pathway that requires Ago catalysis. *Nature* 465, 584-589.

Chen, C.-Y.A., Zheng, D., Xia, Z., and Shyu, A.-B. (2009). Ago–TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. *Nature structural & molecular biology* 16, 1160-1166.

Chen, C.-Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs Modulate Hematopoietic Lineage Differentiation. *Science* 303, 83.

Chen, Y.-W., Song, S., Weng, R., Verma, P., Kugler, J.-M., Buescher, M., Rouam, S., and Cohen, Stephen M. (2014). Systematic Study of Drosophila MicroRNA Functions Using a Collection of Targeted Knockout Mutations. *Developmental cell* 31, 784-800.

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436, 740-744.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460, 479-486.

Choi, W.Y., Giraldez, A.J., and Schier, A.F. (2007). Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. *Science* 318, 271-274.

Chu, C.-y., and Rana, T.M. (2006). Translation Repression in Human Cells by MicroRNA-Induced Gene Silencing Requires RCK/p54. *PLOS Biology* 4, e210.

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., *et al.* (2010). A Novel miRNA Processing Pathway Independent of Dicer Requires Argonaute2 Catalytic Activity. *Science* 328, 1694.

Cimmino, A., Calin, G.A., Fabbri, M., Iorio, M.V., Ferracin, M., Shimizu, M., Wojcik, S.E., Aqeilan, R.I., Zupo, S., Dono, M., *et al.* (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proceedings of the National Academy of Sciences of the United States of America* 102, 13944.

Clurman, B.E., and Hayward, W.S. (1989). Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: evidence for stage-specific events. *Molecular and Cellular Biology* 9, 2657.

Coller, J., and Parker, R. (2005). General Translational Repression by Activators of mRNA Decapping. *Cell* 122, 875-886.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., *et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819-823.

Cook, P.J., Thomas, R., Kannan, R., de Leon, E.S., Drilon, A., Rosenblum, M.K., Scaltriti, M., Benezra, R., and Ventura, A. (2017). Somatic chromosomal engineering identifies BCAN-NTRK1 as a potent glioma driver and therapeutic target. *Nat Commun* 8, 15987.

Court, D.L., Gan, J., Liang, Y.-H., Shaw, G.X., Tropea, J.E., Costantino, N., Waugh, D.S., and Ji, X. (2013). RNase III: Genetics and Function; Structure and Mechanism. *Annual Review of Genetics* 47, 405-431.

Couturier, C.P., Ayyadury, S., Le, P.U., Nadaf, J., Monlong, J., Riva, G., Allache, R., Baig, S., Yan, X., Bourgey, M., *et al.* (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nature communications* 11, 3406.

Davalos, V., Moutinho, C., Villanueva, A., Boque, R., Silva, P., Carneiro, F., and Esteller, M. (2012). Dynamic epigenetic regulation of the microRNA-200 family mediates epithelial and mesenchymal transitions in human tumorigenesis. *Oncogene* 31, 2062-2074.

Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)* 23, 1846-1847.

de Pontual, L., Yao, E., Callier, P., Faivre, L., Drouin, V., Cariou, S., Van Haeringen, A., Geneviève, D., Goldenberg, A., Oufadem, M., *et al.* (2011). Germline deletion of the miR-17~92 cluster causes skeletal and growth defects in humans. *Nature Genetics* 43, 1026-1030.

Denli, A.M., Tops, B.B.J., Plasterk, R.H.A., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature* 432, 231-235.

Di Leva, G., Garofalo, M., and Croce, C.M. (2014). MicroRNAs in Cancer. *Annual Review of Pathology: Mechanisms of Disease* 9, 287-314.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* 29, 15-21.

Dugas, J.C., Cuellar, T.L., Scholze, A., Ason, B., Ibrahim, A., Emery, B., Zamanian, J.L., Foo, L.C., McManus, M.T., and Barres, B.A. (2010). Dicer1 and miR-219 Are required for normal oligodendrocyte differentiation and myelination. *Neuron* 65, 597-611.

Ebert, Margaret S., and Sharp, Phillip A. (2012). Roles for MicroRNAs in Conferring Robustness to Biological Processes. *Cell* 149, 515-524.

Edmonds, M.D., Boyd, K.L., Moyo, T., Mitra, R., Duszynski, R., Arrate, M.P., Chen, X., Zhao, Z., Blackwell, T.S., Andl, T., *et al.* (2016). MicroRNA-31 initiates lung tumorigenesis and promotes mutant KRAS-driven lung cancer. *The Journal of clinical investigation* 126, 349-364.

Eichhorn, S.W., Guo, H., McGeary, S.E., Rodriguez-Mias, R.A., Shin, C., Baek, D., Hsu, S.H., Ghoshal, K., Villen, J., and Bartel, D.P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell* 56, 104-115.

Eis, P.S., Tam, W., Sun, L., Chadburn, A., Li, Z., Gomez, M.F., Lund, E., and Dahlberg, J.E. (2005). Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proceedings of the National Academy of Sciences of the United States of America* 102, 3627.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* 4, 207-214.

Elkayam, E., Faehnle, C.R., Morales, M., Sun, J., Li, H., and Joshua-Tor, L. (2017). Multivalent Recruitment of Human Argonaute by GW182. *Molecular cell* 67, 646-658.e643.

Emery, B. (2010). Regulation of oligodendrocyte differentiation and myelination. *Science* 330, 779-782.

Encell, L.P., Friedman Ohana, R., Zimmerman, K., Otto, P., Vidugiris, G., Wood, M.G., Los, G.V., McDougall, M.G., Zimprich, C., Karassina, N., *et al.* (2012). Development of a dehalogenase-based protein fusion tag capable of rapid, selective and covalent attachment to customizable ligands. *Curr Chem Genomics* 6, 55-71.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5, 976-989.

Eulalio, A., Huntzinger, E., and Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature structural & molecular biology* 15, 346-353.

Eystathioy, T., Chan, E.K.L., Tenenbaum, S.A., Keene, J.D., Griffith, K., and Fritzler, M.J. (2002). A Phosphorylated Cytoplasmic Autoantigen, GW182, Associates with a Unique Population of Human mRNAs within Novel Cytoplasmic Speckles. *Molecular Biology of the Cell* 13, 1338-1351.

Fabian, M.R., Cieplak, M.K., Frank, F., Morita, M., Green, J., Srikumar, T., Nagar, B., Yamamoto, T., Raught, B., Duchaine, T.F., *et al.* (2011). miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT. *Nature structural & molecular biology* 18, 1211-1217.

Fabian, M.R., Mathonnet, G., Sundermeier, T., Mathys, H., Zipprich, J.T., Svitkin, Y.V., Rivas, F., Jinek, M., Wohlschlegel, J., Doudna, J.A., *et al.* (2009). Mammalian miRNA RISC Recruits CAF1 and PABP to Affect PABP-Dependent Deadenylation. *Molecular cell* 35,

868-880.

Fabian, M.R., and Sonenberg, N. (2012). The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature structural & molecular biology* *19*, 586-593.

Fan, H.B., Chen, L.X., Qu, X.B., Ren, C.L., Wu, X.X., Dong, F.X., Zhang, B.L., Gao, D.S., and Yao, R.Q. (2017). Transplanted miR-219-overexpressing oligodendrocyte precursor cells promoted remyelination and improved functional recovery in a chronic demyelinated model. *Scientific reports* *7*, 41407.

Fang, W., and Bartel, David P. (2015). The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. *Molecular cell* *60*, 131-145.

Farazi, T.A., Juranek, S.A., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development (Cambridge, England)* *135*, 1201.

Feingold, M., Hall, B.D., Lacassie, Y., and Martínez-Frías, M.-L. (1997). Syndrome of microcephaly, facial and hand abnormalities, tracheoesophageal fistula, duodenal atresia, and developmental delay. *American Journal of Medical Genetics* *69*, 245-249.

Flemr, M., and Buhler, M. (2015). Single-Step Generation of Conditional Knockout Mouse Embryonic Stem Cells. *Cell reports* *12*, 709-716.

Flemr, M., Malik, R., Franke, V., Nejepinska, J., Sedlacek, R., Vlahovicek, K., and Svoboda, P. (2013). A Retrotransposon-Driven Dicer Isoform Directs Endogenous Small Interfering RNA Production in Mouse Oocytes. *Cell* *155*, 807-816.

Foshay, K.M., and Gallicano, G.I. (2009). miR-17 family miRNAs are expressed during early mammalian development and regulate stem cell differentiation. *Developmental biology* *326*, 431-443.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* *19*, 92-105.

Fukunaga, R., Han, Bo W., Hung, J.-H., Xu, J., Weng, Z., and Zamore, Phillip D. (2012). Dicer Partner Proteins Tune the Length of Mature miRNAs in Flies and Mammals. *Cell* *151*, 533-546.

Gautier, A., Juillerat, A., Heinis, C., Corrêa, I.R., Kindermann, M., Beauflis, F., and Johnson, K. (2008). An Engineered Protein Tag for Multiprotein Labeling in Living Cells. *Chemistry & Biology* *15*, 128-136.

Gibbons, D.L., Lin, W., Creighton, C.J., Rizvi, Z.H., Gregory, P.A., Goodall, G.J., Thilaganathan, N., Du, L., Zhang, Y., Pertsemliadis, A., *et al.* (2009). Contextual extracellular cues

promote tumor cell EMT and metastasis by regulating miR-200 family expression. *Genes Dev* 23, 2140-2151.

Giraldez, A.J., Cinalli, R.M., Glasner, M.E., Enright, A.J., Thomson, J.M., Baskerville, S., Hammond, S.M., Bartel, D.P., and Schier, A.F. (2005). MicroRNAs regulate brain morphogenesis in zebrafish. *Science* 308, 833-838.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal mRNAs. *Science* 312, 75.

Golden, R.J., Chen, B., Li, T., Braun, J., Manjunath, H., Chen, X., Wu, J., Schmid, V., Chang, T.-C., Kopp, F., *et al.* (2017). An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* 542, 197-202.

Gregory, P.A., Bert, A.G., Paterson, E.L., Barry, S.C., Tsykin, A., Farshid, G., Vadas, M.A., Khew-Goodall, Y., and Goodall, G.J. (2008). The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature cell biology* 10, 593-601.

Gregory, R.I., Chendrimada, T.P., Cooch, N., and Shiekhattar, R. (2005). Human RISC Couples MicroRNA Biogenesis and Posttranscriptional Gene Silencing. *Cell* 123, 631-640.

Grimm, J.B., English, B.P., Chen, J., Slaughter, J.P., Zhang, Z., Revyakin, A., Patel, R., Macklin, J.J., Normanno, D., Singer, R.H., *et al.* (2015). A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nat Methods* 12, 244-250, 243 p following 250.

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

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and Mechanisms Related to RNA Interference Regulate Expression of the Small Temporal RNAs that Control *C. elegans* Developmental Timing. *Cell* 106, 23-34.

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. *Molecular cell* 54, 1042-1054.

Gu, J., Wang, M., Yang, Y., Qiu, D., Zhang, Y., Ma, J., Zhou, Y., Hannon, G.J., and Yu, Y. (2018). GoldCLIP: Gel-omitted Ligation-dependent CLIP. *Genomics, Proteomics & Bioinformatics* 16, 136-143.

- Gu, S., Jin, L., Zhang, F., Huang, Y., Grimm, D., Rossi, J.J., and Kay, M.A. (2011). Thermodynamic stability of small hairpin RNAs highly influences the loading process of different mammalian Argonautes. *Proceedings of the National Academy of Sciences* 108, 9208.
- Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835-840.
- Ha, I., Wightman, B., and Ruvkun, G. (1996). A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient formation. *Genes & Development* 10, 3041-3050.
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., *et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129-141.
- Hafner, M., Lianoglou, S., Tuschl, T., and Betel, D. (2012). Genome-wide identification of miRNA targets by PAR-CLIP. *Methods (San Diego, Calif)* 58, 94-105.
- Han, J., Lee, Y., Yeom, K.-H., Kim, Y.-K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development* 18, 3016-3027.
- Han, J., Lee, Y., Yeom, K.-H., Nam, J.-W., Heo, I., Rhee, J.-K., Sohn, S.Y., Cho, Y., Zhang, B.-T., and Kim, V.N. (2006). Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell* 125, 887-901.
- Han, Y.C., Vidigal, J.A., Mu, P., Yao, E., Singh, I., Gonzalez, A.J., Concepcion, C.P., Bonetti, C., Ogradowski, P., Carver, B., *et al.* (2015). An allelic series of miR-17 approximately 92-mutant mice uncovers functional specialization and cooperation among members of a microRNA polycistron. *Nat Genet* 47, 766-775.
- Hauptmann, J., Schraivogel, D., Bruckmann, A., Manickavel, S., Jakob, L., Eichner, N., Pfaff, J., Urban, M., Sprunck, S., Hafner, M., *et al.* (2015). Biochemical isolation of Argonaute protein complexes by Ago-APP. *Proc Natl Acad Sci U S A* 112, 11841-11845.
- Hausser, J., Syed, A.P., Bilen, B., and Zavolan, M. (2013). Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome research* 23, 604-615.
- He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J., *et al.* (2005). A microRNA polycistron as a potential human oncogene. *Nature* 435, 828-833.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcrip-

tion factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654-665.

Hill, D.A., Ivanovich, J., Priest, J.R., Gurnett, C.A., Dehner, L.P., Desruisseau, D., Jarzembowski, J.A., Wikenheiser-Brokamp, K.A., Suarez, B.K., Whelan, A.J., *et al.* (2009). DICER1 Mutations in Familial Pleuropulmonary Blastoma. *Science* **325**, 965.

Hsin, J.P., Lu, Y., Loeb, G.B., Leslie, C.S., and Rudensky, A.Y. (2018). The effect of cellular context on miR-155-mediated gene regulation in four major immune cell types. *Nature immunology* **19**, 1137-1145.

Humphreys, D.T., Westman, B.J., Martin, D.I.K., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16961.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell* **143**, 1174-1189.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Bálint, É., Tuschl, T., and Zamore, P.D. (2001). A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA. *Science* **293**, 834.

Isakova, A., Fehlmann, T., Keller, A., and Quake, S.R. (2020). A mouse tissue atlas of small noncoding RNA. *Proceedings of the National Academy of Sciences* **117**, 25634.

Jackson, E.L., Willis, N., Mercer, K., Bronson, R.T., Crowley, D., Montoya, R., Jacks, T., and Tuveson, D.A. (2001). Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev* **15**, 3243-3248.

Jakymiw, A., Lian, S., Eystathioy, T., Li, S., Satoh, M., Hamel, J.C., Fritzler, M.J., and Chan, E.K.L. (2005). Disruption of GW bodies impairs mammalian RNA interference. *Nature cell biology* **7**, 1267-1274.

Jee, D., Yang, J.-S., Park, S.-M., Farmer, D.J.T., Wen, J., Chou, T., Chow, A., McManus, M.T., Kharas, M.G., and Lai, E.C. (2018). Dual Strategies for Argonaute2-Mediated Biogenesis of Erythroid miRNAs Underlie Conserved Requirements for Slicing in Mammals. *Molecular cell* **69**, 265-278.e266.

Johnston, R.J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845-849.

- Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. (2006). MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS. *Annual Review of Plant Biology* 57, 19-53.
- Jungkamp, A.-C., Stoeckius, M., Mecnas, D., Grün, D., Mastrobuoni, G., Kempa, S., and Rajewsky, N. (2011). In Vivo and Transcriptome-wide Identification of RNA Binding Protein Target Sites. *Molecular cell* 44, 828-840.
- Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev* 19, 489-501.
- Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J.A.F., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nature cell biology* 12, 1014-1020.
- Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics (Oxford, England)* 26, 2204-2207.
- Keppler, A., Gendreizig, S., Gronemeyer, T., Pick, H., Vogel, H., and Johnsson, K. (2003). A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nature biotechnology* 21, 86-89.
- Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell* 115, 209-216.
- Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* 37, 907-915.
- Kim, S.-Y., and Volsky, D.J. (2005). PAGE: parametric analysis of gene set enrichment. In *BMC Bioinformatics*, pp. 144.
- Kim, Y., Yeo, J., Lee, Jung H., Cho, J., Seo, D., Kim, J.-S., and Kim, V.N. (2014). Deletion of Human tarbp2 Reveals Cellular MicroRNA Targets and Cell-Cycle Function of TRBP. *Cell reports* 9, 1061-1074.
- Kleaveland, B., Shi, C.Y., Stefano, J., and Bartel, D.P. (2018). A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* 174, 350-362.e317.
- Klein, U., Lia, M., Crespo, M., Siegel, R., Shen, Q., Mo, T., Ambesi-Impiombato, A., Califano, A., Migliazza, A., Bhagat, G., *et al.* (2010). The DLEU2/miR-15a/16-1 Cluster Controls B Cell Proliferation and Its Deletion Leads to Chronic Lymphocytic Leukemia. *Cancer cell* 17, 28-40.

- Knight, S.W., and Bass, B.L. (2001). A Role for the RNase III Enzyme DCR-1 in RNA Interference and Germ Line Development in *Caenorhabditis elegans*. *Science* 293, 2269.
- Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 17, 909-915.
- Korkmaz, G., Lopes, R., Ugalde, A.P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nature biotechnology* 34, 192-198.
- Korpal, M., Lee, E.S., Hu, G., and Kang, Y. (2008). The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2. *The Journal of biological chemistry* 283, 14910-14914.
- Koulnis, M., Pop, R., Porpiglia, E., Shearstone, J.R., Hidalgo, D., and Socolovsky, M. (2011). Identification and Analysis of Mouse Erythroid Progenitors using the CD71/TER119 Flow-cytometric Assay. *JoVE*, e2809.
- Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2018). miRBase: from microRNA sequences to function. *Nucleic acids research* 47, D155-D162.
- Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic acids research* 47, D155-D162.
- Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., *et al.* (2005). Combinatorial microRNA target predictions. *Nature Genetics* 37, 495-500.
- Krichevsky, A.M., King, K.S., Donahue, C.P., Khrapko, K., and Kosik, K.S. (2003). A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA* 9, 1274-1281.
- Kumar, M.S., Lu, J., Mercer, K.L., Golub, T.R., and Jacks, T. (2007). Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nature Genetics* 39, 673-677.
- Kumar, M.S., Pester, R.E., Chen, C.Y., Lane, K., Chin, C., Lu, J., Kirsch, D.G., Golub, T.R., and Jacks, T. (2009). Dicer1 functions as a haploinsufficient tumor suppressor. *Genes & Development* 23, 2700-2704.
- Kuzuoğlu-Öztürk, D., Huntzinger, E., Schmidt, S., and Izaurralde, E. (2012). The *Caenorhabditis elegans* GW182 protein AIN-1 interacts with PAB-1 and subunits of the PAN2-PAN3 and CCR4-NOT deadenylase complexes. *Nucleic acids research* 40, 5651-5665.

- Kwon, S.C., Baek, S.C., Choi, Y.-G., Yang, J., Lee, Y.-s., Woo, J.-S., and Kim, V.N. (2019). Molecular Basis for the Single-Nucleotide Precision of Primary microRNA Processing. *Molecular cell* 73, 505-518.e505.
- La Rocca, G., Olejniczak, S.H., Gonzalez, A.J., Briskin, D., Vidigal, J.A., Spraggon, L., DeMatteo, R.G., Radler, M.R., Lindsten, T., Ventura, A., *et al.* (2015). In vivo, Argonaute-bound microRNAs exist predominantly in a reservoir of low molecular weight complexes not associated with mRNA. *Proc Natl Acad Sci U S A* 112, 767-772.
- Ladewig, E., Okamura, K., Flynt, A.S., Westholm, J.O., and Lai, E.C. (2012). Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Research* 22, 1634-1645.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* 294, 853-858.
- Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of Tissue-Specific MicroRNAs from Mouse. *Current Biology* 12, 735-739.
- Lai, E.C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* 30, 363-364.
- Lall, S., Grün, D., Krek, A., Chen, K., Wang, Y.-L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., *et al.* (2006). A Genome-Wide Map of Conserved MicroRNA Targets in *C. elegans*. *Current Biology* 16, 460-471.
- Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., *et al.* (2007a). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129, 1401-1414.
- Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., *et al.* (2007b). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129, 1401-1414.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-359.
- Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843-854.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., *et al.* (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425, 415-419.
- Lee, Y., Jeon, K., Lee, J.-T., Kim, S., and Kim, V.N. (2002). MicroRNA maturation: stepwise

processing and subcellular localization. *The EMBO journal* 21, 4663-4670.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* 120, 15-20.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. *Cell* 115, 787-798.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25, 2078-2079.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5, 1752-1779.

Lianoglou, S., Garg, V., Yang, J.L., Leslie, C.S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* 27, 2380-2396.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* 30, 923-930.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 433, 769-773.

Liu, C.-G., Calin, G.A., Meloon, B., Gamliel, N., Sevignani, C., Ferracin, M., Dumitru, C.D., Shimizu, M., Zupo, S., Dono, M., *et al.* (2004a). An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proceedings of the National Academy of Sciences of the United States of America* 101, 9740.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004b). Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305, 1437-1441.

Liu, J., Rivas, F.V., Wohlschlegel, J., Yates, J.R., Parker, R., and Hannon, G.J. (2005a). A role for the P-body component GW182 in microRNA function. *Nature cell biology* 7, 1261-1266.

Liu, J., Valencia-Sanchez, M.A., Hannon, G.J., and Parker, R. (2005b). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nature cell biology* 7, 719-723.

- Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., and Rudensky, A.Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol Cell* **48**, 760-770.
- Logan, M., Pagán-Westphal, S.M., Smith, D.M., Paganessi, L., and Tabin, C.J. (1998). The Transcription Factor Pitx2 Mediates Situs-Specific Morphogenesis in Response to Left-Right Asymmetric Signals. *Cell* **94**, 307-317.
- Los, G.V., Encell, L.P., McDougall, M.G., Hartzell, D.D., Karassina, N., Zimprich, C., Wood, M.G., Learish, R., Ohana, R.F., Urh, M., *et al.* (2008). HaloTag: a novel protein labeling technology for cell imaging and protein analysis. *ACS Chem Biol* **3**, 373-382.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550.
- Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., *et al.* (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-838.
- Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear Export of MicroRNA Precursors. *Science* **303**, 95.
- Maddalo, D., Machado, E., Concepcion, C.P., Bonetti, C., Vidigal, J.A., Han, Y.C., Ogrodowski, P., Crippa, A., Rekhman, N., de Stanchina, E., *et al.* (2014). In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516**, 423-427.
- Makeyev, E.V., Zhang, J., Carrasco, M.A., and Maniatis, T. (2007). The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27**, 435-448.
- Malmeyvik, J., Petri, R., Klussendorf, T., Knauff, P., Åkerblom, M., Johansson, J., Soneji, S., and Jakobsson, J. (2015). Identification of the miRNA targetome in hippocampal neurons using RIP-seq. *Scientific reports* **5**, 12609.
- Marino, S., Vooijs, M., van Der Gulden, H., Jonkers, J., and Berns, A. (2000). Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. *Genes Dev* **14**, 994-1004.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*; Vol 17, No 1: Next Generation Sequencing Data AnalysisDO - 1014806/ej171200.
- Mathonnet, G., Fabian, M.R., Svitkin, Y.V., Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W.C., Darzynkiewicz, E., Pillai, R.S., *et al.* (2007). MicroRNA Inhibition of Translation

Initiation in Vitro by Targeting the Cap-Binding Complex eIF4F. *Science* 317, 1764.

McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P., Wühr, M., Huttlin, E.L., Erickson, B.K., Rad, R., Haas, W., and Gygi, S.P. (2014). MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Analytical Chemistry* 86, 7150-7158.

Medeiros, L.A., Dennis, L.M., Gill, M.E., Houbaviy, H., Markoulaki, S., Fu, D., White, A.C., Kirak, O., Sharp, P.A., Page, D.C., *et al.* (2011). Mir-290–295 deficiency in mice results in partially penetrant embryonic lethality and germ cell defects. *Proceedings of the National Academy of Sciences* 108, 14163.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 Mediates RNA Cleavage Targeted by miRNAs and siRNAs. *Molecular cell* 15, 185-197.

Meister, G., Landthaler, M., Peters, L., Chen, P.Y., Urlaub, H., Lührmann, R., and Tuschl, T. (2005). Identification of Novel Argonaute-Associated Proteins. *Current Biology* 15, 2149-2155.

Meno, C., Gritsman, K., Ohishi, S., Ohfuji, Y., Heckscher, E., Mochida, K., Shimono, A., Kondoh, H., Talbot, W.S., Robertson, E.J., *et al.* (1999). Mouse Lefty2 and zebrafish antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. *Mol Cell* 4, 287-298.

Meno, C., Ito, Y., Saijoh, Y., Matsuda, Y., Tashiro, K., Kuhara, S., and Hamada, H. (1997). Two closely-related left-right asymmetrically expressed genes, lefty-1 and lefty-2: their distinct expression domains, chromosomal linkage and direct neuralizing activity in *Xenopus* embryos. *Genes to Cells* 2, 513-524.

Meno, C., Saijoh, Y., Fujii, H., Ikeda, M., Yokoyama, T., Yokoyama, M., Toyoda, Y., and Hamada, H. (1996). Left–right asymmetric expression of the TGF β -family member lefty in mouse embryos. *Nature* 381, 151-155.

Meno, C., Shimono, A., Saijoh, Y., Yashiro, K., Mochida, K., Ohishi, S., Noji, S., Kondoh, H., and Hamada, H. (1998). lefty-1 is required for left-right determination as a regulator of lefty-2 and nodal. *Cell* 94, 287-297.

Meno, C., Takeuchi, J., Sakuma, R., Koshiba-Takeuchi, K., Ohishi, S., Saijoh, Y., Miyazaki, J.-i., ten Dijke, P., Ogura, T., and Hamada, H. (2001). Diffusion of Nodal Signaling Activity in the Absence of the Feedback Inhibitor Lefty2. *Developmental cell* 1, 127-138.

Miska, E.A., Alvarez-Saavedra, E., Abbott, A.L., Lau, N.C., Hellman, A.B., McGonagle, S.M., Bartel, D.P., Ambros, V.R., and Horvitz, H.R. (2007). Most *Caenorhabditis elegans*

microRNAs Are Individually Not Essential for Development or Viability. *PLOS Genetics* 3, e215.

Moore, M.J., Scheel, T.K., Luna, J.M., Park, C.Y., Fak, J.J., Nishiuchi, E., Rice, C.M., and Darnell, R.B. (2015). miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nat Commun* 6, 8864.

Moore, M.J., Zhang, C., Gantman, E.C., Mele, A., Darnell, J.C., and Darnell, R.B. (2014). Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat Protoc* 9, 263-293.

Mu, P., Han, Y.-C., Betel, D., Yao, E., Squatrito, M., Ogradowski, P., de Stanchina, E., D'Andrea, A., Sander, C., and Ventura, A. (2009). Genetic dissection of the miR-17~92 cluster of microRNAs in Myc-induced B-cell lymphomas. *Genes & Development* 23, 2806-2811.

Mukherji, S., Ebert, M.S., Zheng, G.X.Y., Tsang, J.S., Sharp, P.A., and van Oudenaarden, A. (2011). MicroRNAs can generate thresholds in target gene expression. *Nature Genetics* 43, 854-859.

Müller, M., Fazi, F., and Ciaudo, C. (2020). Argonaute Proteins: From Structure to Function in Development and Pathological Cell Fate Determination. *Front Cell Dev Biol* 7, 360-360.

Nguyen, Tuan A., Jo, Myung H., Choi, Y.-G., Park, J., Kwon, S.C., Hohng, S., Kim, V.N., and Woo, J.-S. (2015). Functional Anatomy of the Human Microprocessor. *Cell* 161, 1374-1387.

Nielsen, C.B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C.B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* 13, 1894-1910.

O'Carroll, D., Mecklenbrauker, I., Das, P.P., Santana, A., Koenig, U., Enright, A.J., Miska, E.A., and Tarakhovsky, A. (2007). A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. *Genes Dev* 21, 1999-2004.

O'Donnell, K.A., Wentzel, E.A., Zeller, K.I., Dang, C.V., and Mendell, J.T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* 435, 839-843.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in *Drosophila*. *Cell* 130, 89-100.

Olejniczak, S.H., La Rocca, G., Gruber, J.J., and Thompson, C.B. (2013). Long-lived microRNA-Argonaute complexes in quiescent cells can be activated to regulate mitogenic responses. *Proc Natl Acad Sci U S A* 110, 157-162.

- Olive, V., Bennett, M.J., Walker, J.C., Ma, C., Jiang, I., Cordon-Cardo, C., Li, Q.-J., Lowe, S.W., Hannon, G.J., and He, L. (2009). miR-19 is a key oncogenic component of mir-17-92. *Genes & Development* 23, 2839-2849.
- Olsen, P.H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol* 216, 671-680.
- Ota, A., Tagawa, H., Karnan, S., Tsuzuki, S., Karpas, A., Kira, S., Yoshida, Y., and Seto, M. (2004). Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma. *Cancer research* 64, 3087-3095.
- Padgett, R.A., Konarska, M.M., Grabowski, P.J., Hardy, S.F., and Sharp, P.A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science* 225, 898.
- Park, S.M., Gaur, A.B., Lengyel, E., and Peter, M.E. (2008). The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev* 22, 894-907.
- Patrick, D.M., Zhang, C.C., Tao, Y., Yao, H., Qi, X., Schwartz, R.J., Jun-Shen Huang, L., and Olson, E.N. (2010). Defective erythroid differentiation in miR-451 mutant mice mediated by 14-3-3 ζ . *Genes & Development* 24, 1614-1619.
- Paulo, J.A., O'Connell, J.D., Everley, R.A., O'Brien, J., Gygi, M.A., and Gygi, S.P. (2016). Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 *S. cerevisiae* proteins across 10 carbon sources. *Journal of proteomics* 148, 85-93.
- Peng, J., Schwartz, D., Elias, J.E., Thoreen, C.C., Cheng, D., Marsischky, G., Roelofs, J., Finley, D., and Gygi, S.P. (2003). A proteomics approach to understanding protein ubiquitination. *Nature biotechnology* 21, 921-926.
- Petersen, C.P., Bordeleau, M.-E., Pelletier, J., and Sharp, P.A. (2006). Short RNAs Repress Translation after Initiation in Mammalian Cells. *Molecular cell* 21, 533-542.
- Pfaff, J., Hennig, J., Herzog, F., Aebersold, R., Sattler, M., Niessing, D., and Meister, G. (2013). Structural features of Argonaute-GW182 protein interactions. *Proc Natl Acad Sci U S A* 110, E3770-3779.
- Piedra, M.E., Icardo, J.M., Albajar, M., Rodriguez-Rey, J.C., and Ros, M.A. (1998). Pitx2 Participates in the Late Phase of the Pathway Controlling Left-Right Asymmetry. *Cell* 94, 319-324.
- Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L., *et al.* (2019). A single-cell molecular

- map of mouse gastrulation and early organogenesis. *Nature* 566, 490-495.
- Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E., and Filipowicz, W. (2005). Inhibition of Translational Initiation by Let-7 MicroRNA in Human Cells. *Science* 309, 1573.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* 26, 841-842.
- Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dunder, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-se- quencing data analysis. *Nucleic acids research* 44, W160-165.
- Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* 8, 2281-2308.
- Rasmussen, K.D., Simmini, S., Abreu-Goodger, C., Bartonicek, N., Di Giacomo, M., Bil- bao-Cortes, D., Horos, R., Von Lindern, M., Enright, A.J., and O'Carroll, D. (2010). The miR-144/451 locus is required for erythroid homeostasis. *Journal of Experimental Medi- cine* 207, 1351-1358.
- Raver-Shapira, N., Marciano, E., Meiri, E., Spector, Y., Rosenfeld, N., Moskovits, N., Bentwich, Z., and Oren, M. (2007). Transcriptional Activation of miR-34a Contributes to p53-Mediated Apoptosis. *Molecular cell* 26, 731-743.
- Rehwinkel, J.A.N., Behm-Ansmant, I., Gatfield, D., and Izaurralde, E. (2005). A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. *RNA* 11, 1640-1647.
- Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates develop- mental timing in *Caenorhabditis elegans*. *Nature* 403, 901-906.
- Richner, M., Victor, M.B., Liu, Y., Abernathy, D., and Yoo, A.S. (2015). MicroRNA-based conversion of human fibroblasts into striatal medium spiny neurons. *Nature Protocols* 10, 1543-1555.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). lim- ma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43, e47.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. *Nature biotechnology* 29, 24-26.
- Rodriguez, C.I., Buchholz, F., Galloway, J., Sequerra, R., Kasper, J., Ayala, R., Stewart,

- A.F., and Dymecki, S.M. (2000). High-efficiency deleter mice show that FLPe is an alternative to Cre-loxP. *Nat Genet* 25, 139-140.
- Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature* 448, 83-86.
- Ruskin, B., and Green, M.R. (1985). An RNA processing activity that debranches RNA lariats. *Science* 229, 135.
- Ryan, A.K., Blumberg, B., Rodriguez-Esteban, C., Yonei-Tamura, S., Tamura, K., Tsukui, T., de la Peña, J., Sabbagh, W., Greenwald, J., Choe, S., *et al.* (1998). Pitx2 determines left-right asymmetry of internal organs in vertebrates. *Nature* 394, 545-551.
- Saijoh, Y., Adachi, H., Mochida, K., Ohishi, S., Hirao, A., and Hamada, H. (1999). Distinct transcriptional regulatory mechanisms underlie left-right asymmetric expression of lefty-1 and lefty-2. *Genes & Development* 13, 259-269.
- Saijoh, Y., Adachi, H., Sakuma, R., Yeo, C.-Y., Yashiro, K., Watanabe, M., Hashiguchi, H., Mochida, K., Ohishi, S., Kawabata, M., *et al.* (2000). Left-Right Asymmetric Expression of lefty2 and nodal Is Induced by a Signaling Pathway that Includes the Transcription Factor FAST2. *Molecular cell* 5, 35-47.
- Sakai, K., and Miyazaki, J.-i. (1997). A Transgenic Mouse Line That Retains Cre Recombinase Activity in Mature Oocytes Irrespective of the Cre Transgene Transmission. *Biochemical and Biophysical Research Communications* 237, 318-324.
- Salomon, William E., Jolly, Samson M., Moore, Melissa J., Zamore, Phillip D., and Serebrov, V. (2015). Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. *Cell* 162, 84-95.
- Sanuki, R., Onishi, A., Koike, C., Muramatsu, R., Watanabe, S., Muranishi, Y., Irie, S., Uneo, S., Koyasu, T., Matsui, R., *et al.* (2011). miR-124a is required for hippocampal axogenesis and retinal cone survival through Lhx2 suppression. *Nature neuroscience* 14, 1125-1134.
- Sato, H., Shien, K., Tomida, S., Okayasu, K., Suzawa, K., Hashida, S., Torigoe, H., Watanabe, M., Yamamoto, H., Soh, J., *et al.* (2017). Targeting the miR-200c/LIN28B axis in acquired EGFR-TKI resistance non-small cell lung cancer cells harboring EMT features. *Scientific reports* 7, 40847.
- Schirle, N.T., and MacRae, I.J. (2012). The Crystal Structure of Human Argonaute2. *Science* 336, 1037.
- Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. *Science* 346, 608.

- Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell* 115, 199-208.
- Sen, G.L., and Blau, H.M. (2005). Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. *Nature cell biology* 7, 633-636.
- Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., *et al.* (2014). Genome-Scale CRISPR-Cas9 Knock-out Screening in Human Cells. *Science* 343, 84.
- Sheth, U., and Parker, R. (2003). Decapping and Decay of Messenger RNA Occur in Cytoplasmic Processing Bodies. *Science* 300, 805.
- Sheu-Gruttadauria, J., and MacRae, I.J. (2018). Phase Transitions in the Assembly and Function of Human miRISC. *Cell* 173, 946-957.e916.
- Shiohama, A., Sasaki, T., Noda, S., Minoshima, S., and Shimizu, N. (2003). Molecular cloning and expression analysis of a novel gene DGCR8 located in the DiGeorge syndrome chromosomal region. *Biochemical and Biophysical Research Communications* 304, 184-190.
- Si, L., Tian, H., Yue, W., Li, L., Li, S., Gao, C., and Qi, L. (2017). Potential use of microRNA-200c as a prognostic marker in non-small cell lung cancer. *Oncology letters* 14, 4325-4330.
- Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution. *Cell* 123, 1133-1146.
- Stark, K.L., Xu, B., Bagchi, A., Lai, W.-S., Liu, H., Hsu, R., Wan, X., Pavlidis, P., Mills, A.A., Karayiorgou, M., *et al.* (2008). Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nature Genetics* 40, 751-760.
- Su, H., Meng, S., Lu, Y., Trombly, Melanie I., Chen, J., Lin, C., Turk, A., and Wang, X. (2011). Mammalian hyperplastic discs Homolog EDD Regulates miRNA-Mediated Gene Silencing. *Molecular cell* 43, 97-109.
- Sutherland, K.D., Song, J.-Y., Kwon, M.C., Proost, N., Zevenhoven, J., and Berns, A. (2014). Multiple cells-of-origin of mutant K-Ras-induced mouse lung adenocarcinoma. *Proceedings of the National Academy of Sciences* 111, 4952.
- Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J., and van der Oost, J. (2014). The evolutionary journey of Argonaute proteins. *Nature structural & molecular biology* 21, 743-753.

- Tam, W., Ben-Yehuda, D., and Hayward, W.S. (1997). *bic*, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA. *Molecular and Cellular Biology* 17, 1490.
- Tam, W., Hughes, S.H., Hayward, W.S., and Besmer, P. (2002). Avian *bic*, a Gene Isolated from a Common Retroviral Site in Avian Leukosis Virus-Induced Lymphomas That Encodes a Noncoding RNA, Cooperates with *c-myc* in Lymphomagenesis and Erythroleukemogenesis. *Journal of Virology* 76, 4275.
- Tan, C.L., Plotkin, J.L., Veno, M.T., von Schimmelmann, M., Feinberg, P., Mann, S., Handler, A., Kjems, J., Surmeier, D.J., O'Carroll, D., *et al.* (2013). MicroRNA-128 governs neuronal excitability and motor behavior in mice. *Science* 342, 1254-1258.
- Tani, H., and Akimitsu, N. (2012). Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. *RNA biology* 9, 1233-1238.
- Ting, L., Rad, R., Gygi, S.P., and Haas, W. (2011). MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods* 8, 937-940.
- Tomari, Y., Matranga, C., Haley, B., Martinez, N., and Zamore, P.D. (2004). A Protein Sensor for siRNA Asymmetry. *Science* 306, 1377.
- Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302, 1212-1215.
- Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147, 1537-1550.
- Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., *et al.* (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature* 583, 711-719.
- Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundaraman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., *et al.* (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 13, 508-514.
- Van Stry, M., Oguin, T.H., Cheloufi, S., Vogel, P., Watanabe, M., Pillai, M.R., Dash, P., Thomas, P.G., Hannon, G.J., and Bix, M. (2012). Enhanced Susceptibility of Ago1/3 Double-Null Mice to Influenza A Virus Infection. *Journal of Virology* 86, 4151.
- Ventura, A., Young, A.G., Winslow, M.M., Lintault, L., Meissner, A., Erkeland, S.J., Newman, J., Bronson, R.T., Crowley, D., Stone, J.R., *et al.* (2008). Targeted deletion reveals

- essential and overlapping functions of the miR-17 through 92 family of miRNA clusters. *Cell* 132, 875-886.
- Wang, H., Moyano, A.L., Ma, Z., Deng, Y., Lin, Y., Zhao, C., Zhang, L., Jiang, M., He, X., Ma, Z., *et al.* (2017). miR-219 Cooperates with miR-338 in Myelination and Promotes Myelin Repair in the CNS. *Developmental cell* 40, 566-582.e565.
- Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* 343, 80.
- Wang, Y., Yang, F., Gritsenko, M.A., Wang, Y., Clauss, T., Liu, T., Shen, Y., Monroe, M.E., Lopez-Ferrer, D., Reno, T., *et al.* (2011). Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *PROTEOMICS* 11, 2019-2026.
- Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell* 151, 1055-1067.
- Wianny, F., and Zernicka-Goetz, M. (2000). Specific interference with gene function by double-stranded RNA in early mouse development. *Nature cell biology* 2, 70-75.
- Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* 75, 855-862.
- Wilson, D.I., Burn, J., Scambler, P., and Goodship, J. (1993). DiGeorge syndrome: part of CATCH 22. *Journal of Medical Genetics* 30, 852.
- Wilson, Ross C., Tambe, A., Kidwell, Mary A., Noland, Cameron L., Schneider, Catherine P., and Doudna, Jennifer A. (2015). Dicer-TRBP Complex Formation Ensures Accurate Mammalian MicroRNA Biogenesis. *Molecular cell* 57, 397-407.
- Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. *Proceedings of the National Academy of Sciences of the United States of America* 103, 4034.
- Yang, J.-S., Maurin, T., Robine, N., Rasmussen, K.D., Jeffrey, K.L., Chandwani, R., Papatrou, E.P., Sadelain, M., Carroll, D., and Lai, E.C. (2010). Conserved vertebrate *mir-451* provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. *Proceedings of the National Academy of Sciences* 107, 15163.
- Yekta, S., Shih, I.H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of *HOXB8* mRNA. *Science* 304, 594-596.

- Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development* 17, 3011-3016.
- Yoshioka, H., Meno, C., Koshiba, K., Sugihara, M., Itoh, H., Ishimaru, Y., Inoue, T., Ohuchi, H., Semina, E.V., Murray, J.C., *et al.* (1998). Pitx2, a Bicoid-Type Homeobox Gene, Is Involved in a Lefty-Signaling Pathway in Determination of Left-Right Asymmetry. *Cell* 94, 299-305.
- Yu, D., dos Santos, C.O., Zhao, G., Jiang, J., Amigo, J.D., Khandros, E., Dore, L.C., Yao, Y., D'Souza, J., Zhang, Z., *et al.* (2010). miR-451 protects against erythroid oxidant stress by repressing 14-3-3 ζ . *Genes & Development* 24, 1620-1633.
- Zhang, H., Kolb, F.A., Jaskiewicz, L., Westhof, E., and Filipowicz, W. (2004). Single Processing Center Models for Human Dicer and Bacterial RNase III. *Cell* 118, 57-68.
- Zhang, K., Zhang, X., Cai, Z., Zhou, J., Cao, R., Zhao, Y., Chen, Z., Wang, D., Ruan, W., Zhao, Q., *et al.* (2018). A novel class of microRNA-recognition elements that function only within open reading frames. *Nature structural & molecular biology* 25, 1019-1027.
- Zhao, J.J., Gjoerup, O.V., Subramanian, R.R., Cheng, Y., Chen, W., Roberts, T.M., and Hahn, W.C. (2003). Human mammary epithelial cell transformation through the activation of phosphatidylinositol 3-kinase. *Cancer cell* 3, 483-495.
- Zhao, X., He, X., Han, X., Yu, Y., Ye, F., Chen, Y., Hoang, T., Xu, X., Mi, Q.S., Xin, M., *et al.* (2010). MicroRNA-mediated control of oligodendrocyte differentiation. *Neuron* 65, 612-626.
- Zhou, X., Sasaki, H., Lowe, L., Hogan, B.L.M., and Kuehn, M.R. (1993). Nodal is a novel TGF- β -like gene expressed in the mouse node during gastrulation. *Nature* 361, 543-547.