COMPUTATIONAL MODELING OF MICRORNA TARGETING AND CONTEXT SPECIFICITY

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yuheng Lu

May 2018

COMPUTATIONAL MODELING OF MICRORNA TARGETING AND CONTEXT SPECIFICITY

Yuheng Lu, Ph.D.

Cornell University 2018

In this dissertation, I present two studies on miRNA regulation enabled by high-throughput sequencing technologies and computational approaches. In the first study, we attempted to learn a general model for miRNA targeting principles based on AGO CLIP and CLASH data. We used discriminative learning on AGO CLIP and CLASH interactions to train a miRNA target prediction model. Our method combined two SVM classifiers, one to predict miRNA-mRNA duplexes and a second to learn AGO's local sequence preferences and positional bias in 3'UTR isoforms. The duplex SVM model enabled the prediction of non-canonical target sites and more accurately resolved miRNA interactions from AGO CLIP data than previous methods. The binding model was trained using a multi-task strategy to learn context-specific and common AGO sequence preferences. The duplex and common AGO binding models together outperformed existing miRNA target prediction algorithms on held-out binding data. In the second study, we attempted to characterize the context specificity of miRNA-mediated regulation of target mRNAs that are co-expressed across multiple cell types. We explored transcriptome-wide targeting and gene regulation by miR-155, whose activation-induced expression plays important roles in innate and adaptive immunity. Through mapping of miR-155 targets using differential AGO iCLIP, mRNA quantification using RNA-Seq,

and 3'UTR usage analysis using polyadenylation (polyA)-Seq in activated miR-155-sufficient and -deficient macrophages, dendritic cells, T and B lymphocytes, we have identified numerous miR-155 targets with cellular context specificity. While alternative cleavage and polyadenylation (ApA) contributed to differential miR-155 binding in some transcripts, a majority of identical 3'UTR isoforms were also differentially regulated, suggesting ApA-independent and cellular context-dependent miR-155-mediated gene regulation reminiscent of sequence-specific transcription factors. Our study provides a comprehensive map of miR-155's regulatory networks in key immune cell types.

BIOGRAPHICAL SKETCH

Yuheng Lu is currently a PhD candidate in the Weill Cornell Graduate School of Medical Sciences of Cornell University. Since 2012, he has worked under the mentorship of Dr. Christina Leslie in the Computational Biology program at Sloan Kettering Institute. His doctoral research involved computational modeling of general microRNA targeting principles and analyses on cellular context specificity of microRNA regulation. Prior to that, he has received his B.S in biological sciences and computer software from Peking University in Beijing, China in 2011.

ACKNOWLEDGEMENTS

Back in the summer of 2011, I came to United States for the first time to start my graduate study. At the age of 20, I was naive and ignorant and had little idea about work and life in general. Thanks to my advisers and coworkers, I now consider myself overall improved as a person, and here I would like to acknowledge their mentorship and support.

I am very fortunate to have Christina Leslie as my adviser. She has been giving me numerous helpful advice on my research over the past six years. My committee members, Olivier Elemento, Michael Kharas and Andrea Ventura, have always been supportive and it is a pleasure to work with all of them.

I would really like to thank my collaborators in the miR-155 project, Jing-Ping Hsin, Gabriel Loeb and Alexander Rudensky. This project would have been impossible without their hard work in designing and carrying out all the experiments. In addition, their rich knowledge in immunology helped me navigate through the vast data sets and discover the information hidden beneath. Without their advice and encouragement, I would have been overwhelmed by the sheer scale of this project and would have never been able to finish it.

Finally, I would like to thank my fellow lab members, who have always been friendly and helpful, and together they make it pleasant and productive to work in this lab. I would like to particularly thank Yuri Pritykin, with whom I have worked together on multiple projects. His dedication to scientific research never ceases to amaze me. I am also very grateful to Hatice Osmanbeyoglu, who has provided me with tremendous help during my job search, as well as reminding me the importance of planning ahead of time.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Molecular basis of miRNA regulation

MicroRNAs (miRNAs) are 20-24 nt long non-coding RNAs that mediate post-transcriptional regulation of target mRNAs. Since the discovery of *lin-4* miRNA in *C. elegans* in 1993 [1], more than 20,000 miRNAs have been identified across eukaryotic organisms [2]. A large proportion of miRNAs are conserved across species, and they also prefer interactions with conserved mRNA target sites [3]. Functional studies have revealed that miRNAs play crucial regulatory roles in numerous developmental processes and diseases [4].

The biogenesis of miRNAs is a tightly regulated multi-step process. miRNA-encoding genes are first transcribed into primary miRNA (pri-miRNA) transcripts by RNA polymerase II (Pol II). The microprocessor complex formed by Drosha, an RNase III family enzyme, and the DGCR8 protein, cleave the pri-miRNAs into shorter precursor miRNA (pre-miRNA) hairpins in the nucleus. The pre-miRNAs are then transported into cytoplasm under the mediation by the exportin XPO5 and the GTP-binding protein Ran. The pre-miRNAs are then further processed by a complex composed of proteins Dicer, Argonaute and TRBP. Dicer (another RNase III family enzyme) cleaves the loop off the hairpin, and the resulting miRNA duplex is then loaded into Argonaute. In the next step the duplex is unwound and one strand, commonly

known as "mature miRNA", remains bound to Argonaute while the other strand known as "miRNA*" is ejected and subsequently degraded, although in some cases both strands can be functional. Notably, a small number of miRNAs are not generated by the canonical pathway described above. For instance, certain pre-miRNAs are produced from splicing of short introns without being processed by Drosha [5].

Argonaute (AGO) family proteins and AGO-bound mature miRNAs together form RNA-induced silencing complexes (RISCs). Protein crystallography has shown that AGO proteins structures have strikingly high conservation across species, even between archaea and human [6–8]. AGO proteins are characterized by four domains: amino-terminal (N), PAZ, MID (middle) and PIWI [9]. The N domain is involved in loading and unwinding of the miRNA duplex. The PAZ domain forms a "binding pocket" that specifically anchors the 3′ end of miRNA, while the MID domain recognizes the 5′ end. The PIWI domain is structurally similar to RNase H and can function as RNA endonucleases, although not all AGO proteins possess RNA cleavage activity. For instance, mammalian genomes encode four AGO proteins AGO1-4, of which only AGO2 is catalytically active [10]. The retained RNA cleavage activity of AGO2 has been found to assist the Dicer-independent maturation process of certain miRNAs, in which pre-miRNA hairpins are directly loaded into RISC and subsequently cleaved. Examples include the maturation of miR-451 and miR-486 during mammalian erythroid development [11]. The Argonaute proteins were subject to regulation by post-translational protein modification. It has been reported that ubiquitylation of AGO proteins by TRIM-NHL family proteins modulates miRNA regulation in both *C. elegans* [12] and mouse [13]. Recently, it has also been discovered that rapid cycles of AGO2

phosphorylation and dephosporylation maintain the pool of available AGO2 proteins and are essential for the global efficiency of miRNA regulation [14].

RISC mediates repression of target mRNAs via two mechanisms. A miRNA can direct slicing of target mRNA when there is extensive base pairing between miRNA and target site and the miRNA is bound to an AGO protein with RNA endonuclease activity [15, 16]. This mechanism of miRNA-mediated repression is common in plants [17] but rarely happens in animals [4]. An alternative mechanism is dominant in mammalian cells, in which RISC is guided by partial pairing between miRNA and target, and does not slice target mRNAs. Instead, RISC recruits the cofactor protein TNRC6 after associating with target mRNA, which interacts with the polyA-binding protein (PABPC) associated with mRNA's polyA tail and also recruits deadenylase complexes PAN2-PAN3 or CCR4-NOT [18]. Both deadenylase complexes shorten the polyA tail, which results in mRNA destabilization. Moreover, CCR4-NOT complex also reduces translational efficiency by binding with the decapping complex at the 5' end of mRNA. Multiple studies have examined the relative extents of mRNA decay and translational repression by comparing mRNA and protein levels after perturbing the expression of miRNAs [19–21]. In most post-embryonic cells, mRNA destabilization explains the majority (~66-90%) of the changes in mRNA expression mediated by miRNAs [22]. Therefore, the expression changes of target mRNAs after miRNA perturbation are commonly used to estimate the extent of miRNA-mediated regulation. However, a recent study observed that translational repression is the predominant consequence of miRNA-mediated regulation in the early zebrafish embryo [23], which may suggest a switch in the post-transcriptional regulation program during embryonic development. In the rest of this thesis, we will focus on the miRNA-mediated regulation

in mammalian cells involving partial miRNA-target complementarity and repression of target mRNA expression.

1.1.2 Established principles of miRNA targeting

Before it is possible to directly map miRNA-mRNA interactions *in vivo* (we will discuss more details in the next section), most of the knowledge about miRNA targeting came from miRNA perturbation experiments combined with sequence analysis. The majority of miRNA target sites that mediate mRNA repression are within 3′ UTR of transcript, although functional target sites within coding sequences and 5′UTR have also been observed [24]. miRNA target recognition is primarily through the Watson-Crick pairing between miRNA nucleotides 2-7 (known as the "seed" sequence at the 5′ end of the miRNA) and mRNA target sites. Many of miRNA target sites also have additional matches to miRNA nucleotide 8, or an A in the target mRNA across from miRNA nucleotide 1, or both [25]. Moreover, the position-1 A in mRNA is always preferred regardless of the first nucleotide of miRNA, suggesting that it does not form base pair with miRNA. This is supported by the structural biology findings showing that a "pocket" in AGO specifically binds to an A at this position in mRNA [8]. In addition to matches in the seed region, pairing with 3′ nucleotides of miRNA, usually at positions 13-16, can also enhance the stability of miRNA-mRNA interactions [26]. Non-canonical target sites, which lack a contiguous 6-mer match to the seed region, can also mediate repression, although in most cases the extent of regulation is significantly weaker than canonical targets [27–31]. Single-molecule assays have suggested that AGO initially scans transcript for target sites with complementarity to only

nucleotides 2–4 of the miRNA, and the initial transient interaction propagates into a stable association only when there are more complementary bases [32], which may help explain the origin of non-canonical targets.

1.1.3 CLIP-Seq and related assays

Crosslinking followed by immunoprecipitation (CLIP) [33] combined with sequencing enables transcriptome-wide characterization of interactions between RNA-binding proteins (RBPs) and their RNA targets. In the original protocol, high-throughput sequencing of RNA isolated by CLIP (HITS-CLIP) [34], cells are first irradiated with 254 nm UV light, inducing the formation of covalent bonds between the amino acid residue and RNA nucleotide in direct contact. Next, the protein-RNA complex is immunoprecipitated with a specific antibody for the protein of interest. The complex is then subject to stringent washing, which will disrupt non-specific protein-RNA interactions but preserve the direct interactions due to the crosslink. Immunoprecipitated RNAs are then treated with optimized concentration of RNase to generate RBP-protected RNA fragments. The protein is then removed via proteinase K digestion. Adapters are ligated to the 5′ and 3′ end of RNA fragments, and the RNA is then reverse transcribed. The cDNA products are PCR amplified with primers that are complementary to the 5′ and 3′ adapter sequences and then sequenced.

An alternative CLIP protocol is photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) [28]. In PAR-CLIP, cells are treated with a modified nucleoside such as 4-thiouridine or 6-thioguanosine, which can be integrated into the newly transcribed RNAs. The modified RNA bases crosslink more efficiently

with RBPs at 365nm UV light. Moreover, the modified uridine bases will be misread by the reverse transcriptase, causing T to C mutations that can be used to pinpoint the crosslinked sites. PAR-CLIP is restricted to cells in culture that can efficiently take up the modified nucleosides.

In both PAR-CLIP and HITS-CLIP protocols, there is a possibility that reverse transcription may stop at nucleotides crosslinked to the remaining peptide after proteinase K digestion [35]. As a result, truncated cDNAs without 5′ adapter sequences will be produced, which will not be PCR-amplified in the later steps of CLIP library preparation. This issue can be resolved by only ligating the 3′ adapter to the crosslinked RNA fragment before reverse transcription and adding the second adapter afterwards, a strategy that has been utilized by improved CLIP protocols to capture the truncated cDNAs. The individual-nucleotide resolution CLIP (iCLIP) protocol [36] uses a reverse transcription primer containing two inversely oriented adapter sequences separated by a restriction site. The cDNA product is then circularized and cleaved at the restriction site, creating a linear sequence with adapters at both ends. Alternatively, the enhanced CLIP (eCLIP) protocol [37] adds a second single-stranded DNA adapter after reverse transcription and RNA removal to allow PCR amplification, in order to avoid the potential loss of material during the circularization and linearization steps. Besides adding more reads to CLIP libraries, the truncated cDNAs captured by improved CLIP protocol are also able to pinpoint the crosslink sites with their 3′ coordinates.

It must be noted, though, that multiple types of biases are present in CLIP libraries and may distort the data analysis. *In vitro* biochemical studies have revealed variations in the crosslinking efficiency for different nucleotides

and amino acid residues [38]. A comprehensive comparison of experimental protocols has shown that the choice of RNase and the condition to digest the RBP-bound RNA has significant impact on the nucleotide conposition of CLIP reads, due to the distinct preferences of different RNases to cleave after certain bases [39]. During the PCR amplification step present in all CLIP protocols, cDNAs in the CLIP library are unevenly amplified [40]. In order to control for PCR amplification artifacts, it has become a common practice in different CLIP protocols to include a randomized barcode sequence in the adapter or reverse transcription primer [27, 36, 37]. Reads aligned to the same genomic coordinate with identical barcode will be treated as PCR duplicates, while the ones with different barcodes will be identified as unique cDNAs. Another source of bias in CLIP libraries is non-specific immunoprecipitations. A study has shown that a large fraction of RBP binding sites from PAR-CLIP libraries overlap with sites captured by non-specific FLAG-GFP immunoprecipitations, and that the same non-specific binding sites are often present in CLIP libraries for different RBPs [41]. Similar non-specific binding sites are also found in libraries generated by the newer eCLIP protocol [37]. Therefore, proper control libraries are essential for correcting systematic biases in CLIP data sets and enabling more accurate identification of RBP binding sites. Control libraries can be generated using IgG or other non-specific antibodies [42]. "Size-matched input controls", which are pre-immunoprecipitation samples prepared identically to the CLIP libraries, have also been used as alternative control libraries for non-specific background binding [37].

An important application of CLIP is mapping the miRNA targets *in vivo*. Since AGO proteins are essential components of RISCs, AGO binding sites captured by CLIP correspond to the binding sites of miRNAs expressed in

given cell types. The regular CLIP protocol is unable to preserve the pairing between miRNAs and their target RNAs. To address this issue, multiple AGO CLIP variants have been developed, namely CLASH [30], iPAR-CLIP [31], CLEAR-CLIP [43] and ChimP [44]. These protocols all feature the usage of RNA ligase to ligate the interacting miRNA and target RNA fragment while both are cross-linked to AGO protein. Successful ligation will generate chimeric reads containing both the miRNA and mRNA sequences, which can be then computationally decoupled to reveal the miRNA-mRNA interactions. Although the ligation efficiency in current protocols are still relatively low (usually <2% [45]), the ligation-based CLIP variants offer great potential for miRNA research due to the capability of directly capturing the *in vivo* interactions between miRNAs and mRNAs.

1.2 Outline of thesis

This dissertation has two independent chapters that cover the major contributions of my graduate research. In the first chapter, I present our novel machine learning algorithm [46] for predicting miRNA targeting based on AGO CLIP and CLASH datasets. In the second chapter, I present a detailed computational analysis of the cellular context-specificity of miR-155 across four key immune cell types, using differential iCLIP between wild-type and miR-155-deficient primary immune cells isolated from mouse.

1.2.1 Learning the general principles of miRNA targeting

Recent technologies like AGO CLIP sequencing and CLASH enable direct transcriptome-wide identification of AGO binding and miRNA target sites, but the most widely used miRNA target prediction algorithms do not exploit these data. We present a novel model for miRNA target prediction through discriminative learning on transcriptome-wide AGO CLIP and CLASH profiles. Our goal was to learn to accurately predict biochemical miRNA-target site interactions, rather than the extent of regulation, in order to increase the sensitivity of miRNA target prediction and learn physiological targeting rules. As the CLASH protocol captures direct interactions between miRNAs and mRNAs by ligation, it provides a partially labeled training set of miRNA-mRNA interactions including many non-canonical pairings, which we combined with canonical AGO binding sites identified by CLIP. We trained one support vector machine (SVM) classifier to model the miRNA-mRNA duplexes and a second SVM to learn AGO's local sequence preferences in the 3'UTR and positional bias in 3'UTR isoforms. The duplex SVM model enables the prediction of both canonical and non-canonical pairings between miRNA and target sequences and outperforms existing methods for assignment of miRNAs to AGO binding sites. The AGO binding model is trained using a multi-task strategy to distinguish between cell type and protocol specific sequence signals and common AGO sequence preferences. The duplex SVM and common AGO binding SVM together outperform existing target prediction approaches when evaluated on held out interaction data.

1.2.2 Context specificity of miRNA regulation in key immune cell types

Numerous microRNAs and their target mRNAs are co-expressed across diverse cell types. However, it is unclear whether they are regulated in a cellular context-independent or -dependent manner. We sought to address this question through computational and comparative genome-wide molecular analyses of RISC bound mRNAs, using individual-nucleotide resolution CLIP (iCLIP) [36], their 3'UTR usage using PolyA-Seq [47] and miR-155-dependent repression (RNA-Seq) in four key immune cell types – activated macrophages, dendritic cells, B cells, and CD4 T cells – isolated from miR-155-sufficient and deficient mice. The analyses of the resulting datasets revealed notable cellular context-dependent miR-155 targeting and regulation of gene expression. While ApA contributed to differential miR-155 binding for some transcripts, in a larger number of cases, identical 3'UTR isoforms were differentially regulated across cell types. These results suggest ApA-independent and cellular context-dependent miR-155-mediated post-transcriptional regulation of gene expression reminiscent of transcriptional regulation by sequence-specific transcription factors. Furthermore, our study provides comprehensive comparative maps of miR-155 regulatory RNA networks as well as global miRNA-mediated Ago binding and genome-wide 3'UTR usage in key activated immune cell types.

CHAPTER 2

LEARNING TO PREDICT MIRNA-MRNA INTERACTIONS FROM AGO CLIP SEQUENCING AND CLASH DATA

Portions of this chapter first appeared in Lu & Leslie [46] and were written in collaboration with Christina Leslie[1].

2.1 Introduction

Recent high-throughput technologies like AGO CLIP sequencing [27] and CLASH (crosslinking, ligation, and sequencing of miRNA-RNA hybrids [30]) enable direct biochemical identification of AGO binding and miRNA target sites transcriptome-wide. The miRNA field has a strong tradition of computationally leveraging transcriptome-wide data to improve target site prediction, but the leading miRNA target prediction methods today do not exploit these new biochemical data. Here we present a systematic approach to learn both the rules of miRNA-target site pairing and a binding model of AGO's local sequence preferences and positional bias in alternative 3'UTR isoforms in order to accurately predict miRNA-target interactions.

Before it became possible to map AGO-mRNA and miRNA-mRNA interactions directly, the major advance in miRNA target prediction came from restricting to predefined classes of miRNA seed matches in 3'UTRs and training a model to predict mRNA expression changes in miRNA overexpression experiments. TargetScan [26] was the first algorithm to introduce the strategy of

[1] As per the Cornell dissertation guidelines, the dissertation can include material that has been previously published or is soon to be published.

correlating context features of miRNA seed sites-including flanking AU content, position in the 3'UTR, and complementarity to the 3' end of the miRNA-with extent of target down-regulation in miRNA transfection experiments. Similar observations were encapsulated in the TargetRank method [48], and these studies established that rules of miRNA targeting could be statistically decoded from transcriptome-wide data.

However, new data from AGO CLIP sequencing and CLASH challenge some of the assumptions of existing prediction strategies. These data confirm the prevalence of non-canonical target sites lacking complementarity to the miRNA 2-7 (6-mer) seed region and conversely show that even exact miRNA 2-8 (7-mer) seed matches are often not AGO bound [29]. Meanwhile, most target prediction methods require strong seeds to avoid false positives. For example, predictions from TargetScan 7.0 [49] still require either perfect 2–8 seed complementarity (7-mer-m8 site) or a 2–7 seed with A across from miRNA position 1 (7-mer-1A site), although AGO CLIP data suggests that 7-mer and 8-mer seeds are found in only about half of AGO binding sites [29]. The mirSVR method [50], which also trains on miRNA overexpression experiments, allows up to one mismatch or G:U wobble in the 6-mer seed region, but in practice few non-canonical sites are assigned even moderate scores. Therefore, current target prediction methods may focus on detecting the most effective miRNA sites at the cost of missing a large proportion of miRNA-mRNA interactions. Furthermore, training on non-physiological miRNA overexpression experiments may obscure more subtle targeting rules.

A few studies have developed algorithms to resolve which highly expressed miRNAs are associated with individual AGO CLIP peaks. For example,

microMUMMIE is an algorithm for analysis of AGO PAR-CLIP that uses the location of T-to-C mutations—indicative of the site of cross-linking of the RNA-binding protein to the RNA in the PAR-CLIP assay—to assign the most likely canonical seed [51]. Other methods use energy-based duplex prediction to associate miRNAs with CLIP-mapped target sequences. In particular, MIRZA uses an unsupervised probabilistic approach to learn parameters of a duplex alignment model from AGO CLIP peaks, and the duplex model can be used to make *de novo* miRNA target site predictions from 3'UTR sequence [52]. Note that the MIRZA study used the term "non-canonical" to refer to sites lacking 7 or 8 nucleotides of perfect complementarity to the 5' end of the miRNA; therefore, their reported non-canonical sites included both perfect 6-mer and many 7-mer-1A sites. (We will use "non-canonical" exclusively for sites lacking full complementarity in the 2–7 6-mer seed region.) More recently, MIRZA-G combined MIRZA duplex quality scores with known context features like flanking AU content and predicted secondary structure accessibility as well as conservation, once again to predict extent of down-regulation in miRNA overexpression experiments [53].

Here we present a novel model for miRNA target prediction through discriminative learning on transcriptome-wide AGO CLIP and CLASH profiles. Our goal was to learn to accurately predict biochemical miRNA-target site interactions, rather than the extent of regulation, in order to increase the sensitivity of miRNA target prediction and learn physiological targeting rules. As the CLASH protocol captures direct interactions between miRNAs and mRNAs by ligation, it provides a partially labeled training set of miRNA-mRNA interactions including many non-canonical pairings, which we combined with canonical AGO binding sites identified by CLIP. We trained one

support vector machine (SVM) classifier to model the miRNA-mRNA duplexes and a second SVM to learn AGO's local sequence preferences in the UTR and positional bias in 3'UTR isoforms. The duplex SVM model enables the prediction of both canonical and non-canonical pairings between miRNA and target sequences and outperforms existing methods for assignment of miRNAs to AGO binding sites. The AGO binding model is trained using a multi-task strategy to distinguish between cell type and protocol specific sequence signals and common AGO sequence preferences. The duplex SVM and common AGO binding SVM together outperform existing target prediction approaches when evaluated on held out interaction data.

2.2 Methods

2.2.1 Feature representation for duplex and context models

We adapted the feature representation from MIRZA [52] to describe the duplex structures formed between interacting (miRNA, site) pairs. Three types of features were included in the representation: (1) the type of base pair (GU, UG, AU, UA, GC, CG) at each position in the alignment; (2) the bases where a loop is opened, symmetrically extended or asymmetrically extended in the duplex structure; (3) binary variables for each position in the miRNA sequence representing whether it is paired to an mRNA base or not. One major change we made to the original representation was that the only permissible base pairing of the first base in the miRNA was with an A in mRNA sequence, so that only an A across from position 1 would contribute positively to the score. This restriction

is derived from the observations in previous studies [26].

We described the mRNA sites with two types of UTR features: local sequence context and global positional context. The sequence context was represented by positional k-mer features (k = 1, ..., 6) from 30 nt sequences upstream and downstream of the miRNA seed match and implemented using two weighted degree string kernels [54]. Three positional context features for each site were computed as (i) the distance to the nearest stop codon, (ii) the distance to the next end of a 3'UTR isoform, and (iii) the distance to the previous end of a 3'UTR isoform and were renormalized with a radial basis kernel. These local sequence kernel and positional kernel were then combined by summing kernel matrices.

2.2.2 Training and testing of duplex and context models

We trained the duplex model both on (miRNA, site) examples directly derived from CLASH interactions and on examples with interactions inferred from CLIP based on 6-mer seed complementarity. One major advantage of the miRNA-mRNA duplex representation described above is that the model weights w can also be used as the parameters for local pairwise alignment [52]: given the feature description φ(miRNA, site) for a duplex alignment, the alignment score can be described by the additive scoring function $w \cdot \varphi$(miRNA, site). Therefore, by iteratively optimizing the model weights given the currents alignments and then computing the optimal alignments given current model weights, we can simultaneously optimize the duplexes and the scoring model. The initial duplex structure for each (miRNA, site) pair was predicted by

duplexfold in the ViennaRNA package [55], and the corresponding duplex feature vectors were then used to train a linear support vector machine (SVM) classifier. The model weights w were then used as local alignment parameters to update the duplex structure between the miRNA and mRNA site sequences. The same process was repeated for 12 iterations, by which point the model vector had converged, and the final duplex structures and model weights were used as the duplex model's output. To compensate for the class imbalance, in each iteration we only used a fraction of negative examples randomly sampled from the whole set while using all positive examples. Specifically, we sampled 15 times as many CLASH negatives as CLASH positives, and the same number of CLIP negatives as CLIP positives.

We applied a regular SVM classifier to the UTR kernel matrix when we trained the AGO binding model using CLIP training data from a single cell type. When we combined data sets from multiple cell types, we applied the multi-task learning approach [56] and treated the different cell types as different but related learning tasks to address the possibility of cell type specific miRNA targeting and AGO binding rules as well as protocol specific biases. We implemented the multi-task SVM as a modification to the kernel matrix:

$$K_{st}(x, z) = (\mu + \delta_{st})K(x, z)$$

If two examples x and z belong to the same task (in other words, two sites were from the same cell type), then an extra weight is added to their product in the kernel matrix to reflect the relationship. The free parameter μ controls the closeness of task-specific models to the average model, and its optimal value was determined by five-fold cross-validation. All the machine learning procedures described above were implemented with Numpy (`http:`

//www.numpy.org) and the Shogun machine learning tool box (http://www.shogun-toolbox.org).

2.3 Results

2.3.1 ChimiRic learns both miRNA-mRNA duplex structures and AGO binding preferences from CLIP and CLASH data

ChimiRic's duplex model is trained on chimeric reads from CLASH data, which associates a miRNA with a target sequence via chimeric reads and can identify non-canonical binding sites, and AGO CLIP binding sites containing a 6-mer seed match (or longer seed) for a single highly expressed miRNA. In the latter case, differential AGO CLIP-seq analysis suggests that an AGO bound site that can be associated with a unique miRNA by a canonical 6-mer seed is likely a binding site for that miRNA.

We used CLASH [30] and AGO PAR-CLIP data [39] in HEK293 cells to train the duplex model, restricting to the top 59 expressed miRNAs in 21 miRNA seed families. To compile the training set, sites identified by CLASH chimeric reads were required to fall within 3'UTRs, contain a sequence within an edit distance of 1 (substitutions or indels) from a canonical 6-mer seed match for the interacting miRNA, and also be supported by non-chimeric reads. This filtering yielded the *positive* training examples consisting of 1,727 (miRNA, site) pairs supported by chimeric reads, of which 1,228 were non-canonical interactions,

together with 11,211 canonical (miRNA, site) examples from AGO CLIP sites (Figure 2.1a). Canonical miRNA seed matches that are not AGO bound based on CLIP data, together with (miRNA, site) pairs where an AGO-bound site is paired with an incorrect miRNA, provided 25,411 *negative* examples. To compensate for the class imbalance, we only used a randomly sampled subset of negative examples in training.

We trained a structural SVM [57] on positive and negative (miRNA, site) training examples to learn a model for predicting miRNA-site duplex alignments. Here, the model vector w of the SVM represents the scoring parameters for local pairwise alignment. SVM training proceeds iteratively, alternating between obtaining optimal alignments of all training examples given the current SVM parameters w and updating the model vector w given the current duplex alignments. The model update step involves solving the SVM large-margin optimization problem so that the discriminant scores assigned to positive and negative (miRNA, site) examples have the correct sign and obey margin constraints, with a hinge loss function to control margin violations. To define the local alignment scoring system and convert the alignment score into an SVM discriminant function, we used a parameterization similar to the energy-based scoring system in MIRZA, namely a match/mismatch score that depends on the position in the miRNA sequence together with the nucleotides being aligned and penalties for loop opening and for symmetric and asymmetric loop extensions. One important difference with MIRZA is that the chimiRic alignment can only start at position 1 of the miRNA if is it matched against nucleotide A, which more accurately reflects known determinants of miRNA targeting.

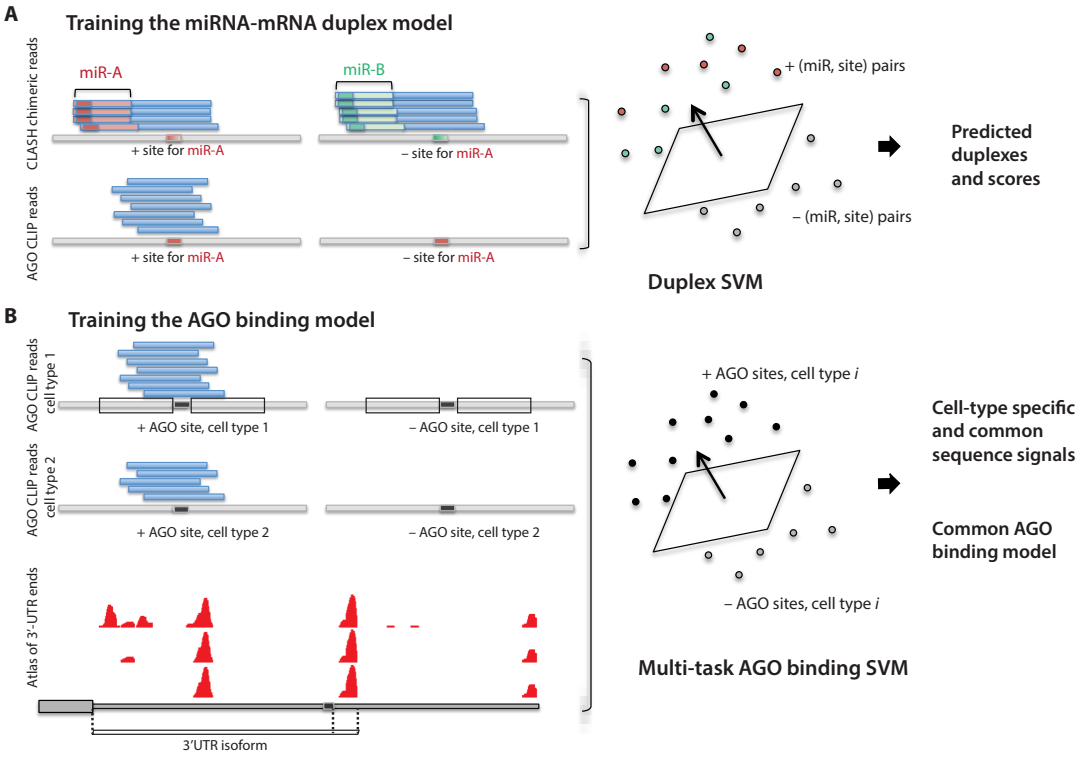


Figure 2.1: **Overview of the chimiRic prediction model.** (A) The first component of the chimiRic model is the duplex SVM, which learns to predict and score miRNA-mRNA duplex alignments from CLASH and CLIP-seq data. Positive (miRNA, site) training examples comprise canonical and non-canonical pairings identified by chimeric reads in CLASH data (top left) as well as sites with canonical miRNA seeds supported by AGO CLIP data (bottom left). Negative (miRNA, site) training examples include sites that are paired with a different miRNA based on CLASH chimeric read data (top right) or miRNA seed matches with no AGO CLIP evidence (bottom right). The duplex SVM learns the parameters for local duplex sequence alignment and predicts optimal alignments for (miRNA, site) pairs through an iterative training procedure. (B) The second component of chimiRic is the AGO binding SVM, which uses features encoding the positional bias of AGO binding sites relative to (possibly multiple) 3′ ends of transcripts as well as the local positional k-mer sequence features. Mouse and human ApA atlases based on 3′ end sequencing data (bottom) provide the coordinates of 3′ ends used in the analysis.

The second component of chimiRic's scoring system is an SVM classifier that learns to discriminate the local sequence features and positional bias in 3′UTR isoforms of true AGO binding sites versus sites that contain 6-mer seed matches

of highly expressed miRNAs but are not AGO-bound, as determined by CLIP data (Figure 2.1b). Here we considered two AGO CLIP sequencing data sets, the human HEK293 PAR-CLIP data set as well as a HITS-CLIP data set in activated mouse CD4 T cells [29]. The local sequence context of the upstream and downstream 30 nt regions flanking the 6-mer seed match are represented using weighted degree kernels [54], which encode position specific k-mers for k = 1 ... 6. The positions of 3′ ends of alternative 3′UTR isoforms were identified from a human 3′-seq tissue atlas [58] and a mouse PolyA-seq atlas [47]. For each site in human or mouse, positional information was encoded by a vector of distance values (measured in nucleotides) to the annotated stop codon and to the nearest mapped 3′ ends and transformed using a radial basis kernel, and the sum of the weighted degree kernels and positional radial basis kernel was used to train the SVM. In order to model differences in AGO binding preferences between the two data sets—both due to protocol differences and potentially due to cell-type specific factors influencing AGO occupancy—we used multi-task learning to train cell-type specific AGO preference models together with a common AGO binding model (Figure 2.1b). The cell-type specific models are intended to absorb sequence signals that predict AGO binding in a context-dependent manner, while the common model can be used for target prediction in any new context.

2.3.2 ChimiRic's duplex model outperforms existing methods for predicting miRNA-mRNA interactions supported by chimeric reads

To evaluate chimiRic's duplex model, we held out from training all HEK293 CLASH interactions for a single miRNA seed family (positive test examples) together with a collection of targets sites that interact with other miRNAs based on chimeric read evidence (negative test examples), and we assessed whether the model could rank the held-out miRNA family's true target sites above these other sites. We found that the duplex model could more accurately discriminate true from false interactions compared to MIRZA [52], an existing method for learning miRNA-mRNA interactions from CLIP data, based on area under the ROC curve (auROC) analysis (Figure 2.2a, blue points, $p <$ 3.02e-5, signed rank test). Note that the original MIRZA model was trained on the same HEK293 PAR-CLIP data set as we used to train the duplex model. To further evaluate the performance on independent data sets, we then used the duplex model trained on HEK293 CLIP and CLASH data to predict miRNA-mRNA interactions supported by chimeric reads from iPAR-CLIP in *C. elegans* [31] and CLEAR-CLIP in mouse brain [43]. Again, chimiRic's duplex model outperformed MIRZA for the task of ranking observed interactions for each miRNA seed family above interactions with targets sites of other miRNAs in both *C. elegans* (Figure 2.2a, green points, $p <$ 1.45e-2, signed rank test) and mouse brain (Figure 2.2a, purple points, $p <$ 4.87e-2, signed rank test) data sets. These results suggest that chimiRic's miRNA-mRNA duplex model can generalize across organisms and protocols for mapping miRNA-mRNA interactions.

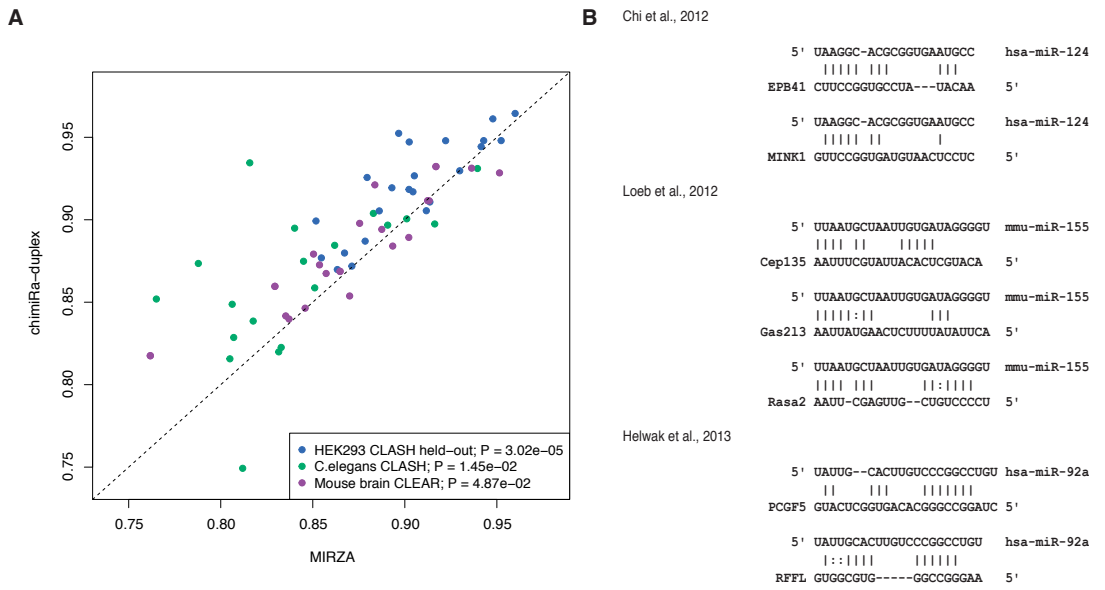


Figure 2.2: **Performance of chimiRic's duplex model for predicting miRNA-mRNA interactions supported by chimeric reads.** (A) Duplex model's performance for predicting the correct interacting miRNA seed family among miRNA-mRNA interactions supported by CLASH chimeric reads. For each miRNA seed family tested, all CLASH-supported interactions for miRNAs in the family are held out from training and form the positive test set; negative test examples consist of interactions for a collection of miRNAs that are held out from training in all experiments. Each point represents the held-out auROC for one of the top 23 miRNA seed families in HEK293 (blue), top 19 miRNA seed families in C. elegans (green) and top 20 miRNA seed families in mouse brain (purple). (B) Examples of duplexes predicted by the model for previously validated non-canonical miRNA-mRNA interactions. Various non-canonical miRNA-mRNA interaction modes were represented, including GU wobbles, bulges and mismatches within seed sequences and interactions relying on 3' base pairing instead of seed pairing.

Previous differential CLIP and CLASH studies have revealed a broad spectrum of non-canonical miRNA-mRNA interaction modes, including GU wobbles, bulges and mismatches within seed sequences, and interactions relying on 3' base pairing instead of seed pairing [29, 30, 59]. In order to test whether our duplex model captures some of these known patterns of non-canonical binding, we predicted duplexes for a variety of non-canonical

miRNA target sites that have been validated by luciferase assays in previous studies (Figure 2.2b). Our model have not only correctly identified the correct interacting miRNA above the other highly expressed miRNAs, despite the lack of exact 6-mer seed matches, but also produced duplex structures representative of the previously described interaction modes, including GU wobbles, mismatches and bulges in the seed region, and complementary base pairings in the 3' region (Figure 2.2b).

2.3.3 The full chimiRic model outperforms traditional target prediction for discriminating CLIP-supported miRNA binding sites

Next we combined the duplex model with the AGO binding model, which is trained to discriminate between true AGO bound sites containing 6-mer seeds for highly expressed miRNAs and sites with 6-mer seeds that are not supported by AGO CLIP read evidence, based both on local sequence context and positional bias within 3'UTR isoforms. We used a multi-task strategy to train on AGO-bound versus unbound canonical seed sites for highly expressed miRNAs in two AGO CLIP data sets, HEK293 PAR-CLIP [39] and HITS-CLIP in mouse CD4 T cells [29]. This procedure learned both task-specific SVM models of AGO binding and a common SVM model. The task-specific SVMs may capture protocol-specific CLIP biases and/or cell-type specific AGO binding preferences. For target prediction in a new context where no CLIP data is available, the common SVM provides a "cell-type agnostic" model of AGO sequence and position preferences.

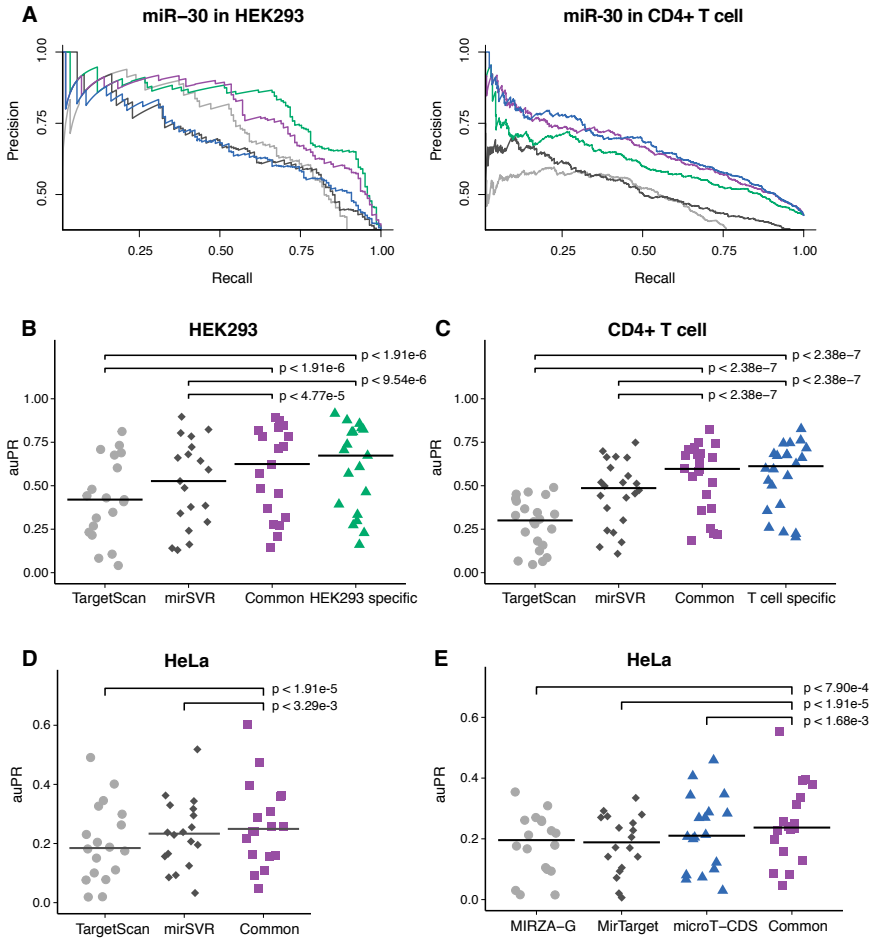


Figure 2.3: **Performance comparison between chimiRic and other methods for discriminating AGO bound sites from unbound sites.** (A) Examples of precision-recall curves for discriminating AGO-bound canonical target sites from seeds with no AGO support for a single miRNA family (miR-30) in HEK293 and CD4 T cell. Curves correspond to task-specific (T cell: blue; HEK293: green) and common (purple) AGO binding models, TargetScan (grey) and mirSVR (black). (B, C) Performance of TargetScan, mirSVR and task-specific/common AGO binding models on held-out miRNA families in HEK293 and CD4 T cells measured by auPR. Crossbars represent the median auPR of each model. (D) Performance of TargetScan, mirSVR and the common AGO binding model on the top miRNA families in an independent HeLa CLIP-seq data set measured by auPR. Crossbars represent the median auPR of each model. (E) Performance of MIRZA-G (grey), MirTarget (black), DIANA-microT-CDS (blue) and the common AGO binding model (purple) on the top miRNA families in an independent HeLa CLIP-seq data set measured by auPR. Crossbars represent the median auPR for each model.

To evaluate the combined chimiRic model, for each miRNA seed family, we held out all HEK293 positive target site sequences—both canonical and non-canonical sites supported by chimeric reads from CLASH as well as canonical sites with AGO CLIP read evidence that can be unambiguously assigned to the seed family—and negative site sequences, for training both the duplex and AGO binding models. We then asked how well the combined model performs at discriminating AGO-bound from unbound canonical sites relative to TargetScan [49] and mirSVR [50], two widely used miRNA target prediction algorithms. Figure 2.3a shows precision-recall curves for the combined chimiRic duplex and HEK293-specific AGO binding model as well as for TargetScan and mirSVR for prediction of canonical sites for several miRNA families. Since TargetScan requires greater seed complementarity than the canonical 6-mer seed (either 7-mer 1A or complementary at miRNA positions 2–8), its overall recall of biochemically-defined sites is limited (note that while the TargetScan 7.0 release discusses 6-mer seeds and non-canonical seeds [49], only a very small fraction of sites were non-canonical in the prediction download files). Evaluating performance by area under the precision-recall curve (auPR) across held-out miRNA seed families showed that this performance advantage was significant over TargetScan (Figure 2.3b, $p < 1.91\text{e-}6$, signed rank test) and mirSVR (Figure 2.3b, $p < 9.54\text{e-}6$, signed rank test). Moreover, even measuring performance up to 50% recall (auPR50), where there are still AGO-bound 7-mer sites to detect, chimiRic still outperformed TargetScan on held-out miRNAs in the HEK293 and CD4 T cell data sets. We then tested the combination of chimiRic's duplex model and the common AGO binding model. Again we found that chimiRic significantly out-performed TargetScan (Figure 2.3b, $p < 1.91\text{e-}6$, signed rank test) and mirSVR (Figure 2.3b, $p < 4.77\text{e-}5$, signed rank

test) on held-out miRNA seed families in HEK293, with minor difference in chimiRic's performance compared to the HEK293-specific model. Similarly, when predicting the biochemically defined target sites of held-out miRNA families in CD4 T cells, chimiRic's duplex model combined with either the T cell specific or the common AGO binding model outperformed TargetScan on held-out miRNAs in the HEK293 and T cell data sets. We then tested the combination of chimiRic's duplex model and the common AGO binding model. Again we found that chimiRic significantly outperformed TargetScan (Figure 2.3b, $p < 1.91e\text{-}6$, signed rank test) and mirSVR (Figure 2.3b, $p < 4.77e\text{-}5$, signed rank test) on held-out miRNA seed families in HEK293, with minordifference in chimiRic's performance compared to the HEK293-specific model. Similarly, when predicting the biochemically defined target sites of held-out miRNA families in CD4 T cells, chimiRic's duplex model combined with either the T cell specific or the common AGO binding model outperformed TargetScan (Figure 2.3c, $p < 2.38e\text{-}7$ and $p < 2.38e\text{-}7$, signed rank tests) and mirSVR (Figure 2.3c, $p < 2.38e\text{-}7$ and $p < 2.38e\text{-}7$, signed rank tests).

As an independent validation, we also evaluated chimiRic's performance in a third cellular context using two HITS-CLIP data sets in HeLa cells [27, 59]. Again, we found that the common AGO binding model combined with duplex model had a significant advantage over TargetScan (Figure 2.3d, $p < 1.91e\text{-}5$, signed rank test) and mirSVR (Figure 2.3d, $p < 3.29e\text{-}3$, signed rank test). Evaluation using auPR50, which favors TargetScan by allowing reduced recall, still showed a significant performance advantage of the common chimiRic model over TargetScan and mirSVR in HEK293 and T cells, with a statistical tie on the HeLa cells. We also evaluated the performance of three additional methods, MIRZA-G [53], MirTarget [60] and DIANA-microT-CDS [61], all of

which are trained on AGO CLIP data and provide one a single prediction score for each miRNA-gene interaction. When we compared the performance on the same HeLa data set, the common chimiRic model outperformed all three methods measured by auPR (Figure 2.3e, $p < 7.90\text{e-}4$, $p < 1.91\text{e-}5$ and $p < 1.68\text{e-}3$, signed rank test), partly due to chimiRic's better recall. When measured by auPR50, chimiRic still achieved a statistical tie against these methods, showing that chimiRic's top-ranked predictions are at least as accurate as other methods trained on AGO CLIP data sets.

2.3.4 AGO-binding model learns 3'UTR positional preferences and RNA-binding motifs associated with miRNA targeting

Previous studies have suggested that 3'UTR miRNA target sites tend to reside near the stop codons or near the 3' end of the transcript rather than the middle of 3'UTRs [26]. We confirmed a positional enrichment of AGO-bound sites near the stop codons (Figure 2.4a, top) and near the end of the 3'UTR compared to miRNA seeds with no AGO binding in CD4 T cells across mouse transcripts. Additionally, for multi-UTR transcripts, we observed an enrichment of AGO-bound sites in the region upstream of internal 3' cleavage sites (as mapped by PolyA-seq) that was absent for the negative site examples (Figure 2.4a, top, $p < 2.2\text{e-}16$, KS test). We also observed an enrichment of positive site examples ~200nt downstream of internal cleavage sites, suggesting that the resolution of the mapped 3' ends in the mouse atlas is limited and/or that clusters of nearby 3' cleavage sites confound the analysis.

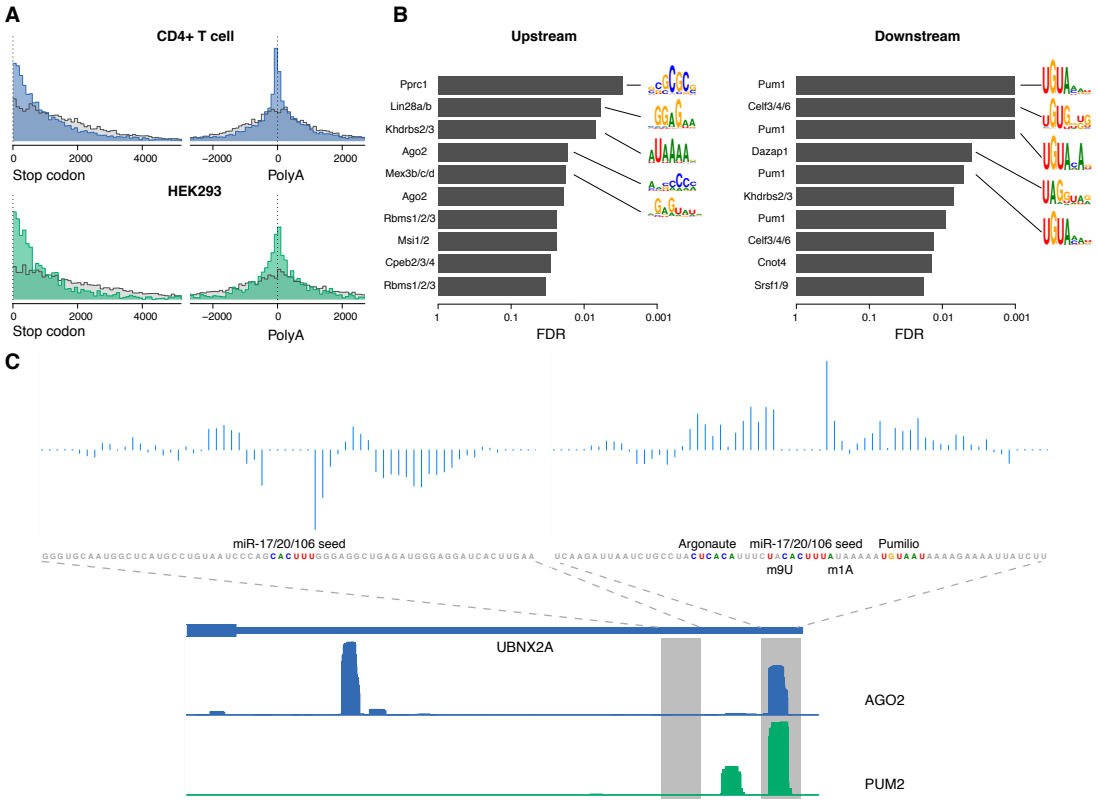
To further interpret the sequence features in the AGO binding model,

we used the positional oligomer importance matrix (POIM) [62] approach to identify the significant positional k-mers. From the 1-mer POIMs, we observed not only high AU content flanking the miRNA seed matches in general but also specific positional signals like m1A and m8/9U, which are consistent with findings from previous studies [48]. Moreover, the representation allowed us to go beyond single nucleotide composition, which is the extent of sequence contextual information used in most previous miRNA target prediction methods, to explore more complex sequence features.

Previous studies have suggested that various RNA binding proteins (RBPs) can bind to regions proximal to miRNA target sites in order to enhance or repress miRNA-mediated regulation [63–65]. Therefore, one potential explanation for the long positional k-mers that discriminate between AGO binding sequences and unbound sequences is that they correspond to the motifs of co-binding RBPs that mediate AGO occupancy. To explore this hypothesis, we matched the 6-mers from positions with top differential POIM scores to RNAcompete *in vitro* affinity data for a compendium of RBPs [66, 67]. By measuring the enrichment of these k-mers in RNAcompete data across all RBPs and assessing significance relative to an empirical null model based on training SVMs on random permutations of the class labels, we found that the position-specific k-mers in upstream and downstream sequences were indeed consistent with several known RBP motifs (Figure 2.4b). In the common AGO-binding model, we identified an AC-rich motif upstream of the seed match that matched an AGO RNAcompete experiment and has been proposed to be the miRNA-independent binding signal for Argonaute [67]. Meanwhile, in the downstream component of the common model, Pumilio was identified as the most significant RBP motif. It has been previously suggested that Pumilio

has a role in regulating miRNA site accessibility of specific target genes [64, 65, 68]. Our analysis suggests that Pumilio may play a transcriptome-wide role in mediating AGO binding. We compared the HEK293 AGO CLIP to PUM2 PAR-CLIP in the same cell type and found that 16.4% of AGO sites in HEK293 overlapped with PUM2 binding sites. Figure 2.4c shows one example of a miR-17/20/106 target site in the 3'UTR of UBNX2A together with sequence signals identified by the model. After decomposing the SVM sequence scores into positional prediction scores, we found that the positions with positive contribution overlapped exactly with the Pumilio binding motif and Pumilio CLIP coverage. In contrast, another miR-17/20/106 seed match site in the same 3'UTR was not bound by AGO and lacked significant positional k-mers from the sequence model.

Figure 2.4: **Interpretation of the AGO-binding model learned from CLIP-seq data.** (A) Positional distribution of AGO binding sites (blue/green) and unbound sites (grey) within 3'UTRs in CD4 T cell (top) and HEK293 (bottom), showing enrichment of bound sites near the start of the 3'UTR (left) and in the region upstream of internal 3' cleavage sites of multi-UTR transcripts (right). There is also enrichment of AGO-bound sites ~200nt downstream of internal 3' cleavage sites, suggesting that the resolution of the PolyA-seq peaks can be limited and/or that clusters of nearby 3' cleavage sites confound the analysis. All distances were between the position aligned against nucleotide 2 of the miRNA and the start/end of the corresponding 3'UTR. (B) RBPs with motifs that match the most discriminative k-mers in the common sequence model. Positions with the highest differential POIM for 6-mers upstream and downstream of the miRNA seeds were chosen, and then a signed rank test was used to assess the enrichment of POIM k-mers in RNAcompete array probes. False discovery rates (FDRs) were estimated using the empirical p-value distribution from 1,000 SVMs trained on random permutations of the +/- labels. Motif logos summarized from the original RNAcompete assays are shown for the top 5 RBPs as ranked by FDR. (The same RBP symbol may appear multiple times since in some cases several constructs of the same protein were assayed by RNAcompete.) (C) An example of co-binding of Pumilio and Argonaute at miRNA target sites. Two miR-17/20/106 seed matches within the 3'UTR of UBNX2A are shown, one with AGO2 binding and one without, along with the coverage profiles of AGO2 and PUM2 CLIP in HEK293. For each site, the prediction scores from the SVM sequence model are decomposed into positional scores and displayed. Sequence features near the target site including the Argonaute motif, Pumilio motif, m1A and m9U are also highlighted.


A
CD4+ T cell
Stop codon
PolyA
HEK293
Stop codon
PolyA
B
Upstream
Pprc1
Lin28a/b
Khdrbs2/3
Ago2
Mex3b/c/d
Ago2
Rbms1/2/3
Msi1/2
Cpeb2/3/4
Rbms1/2/3
FDR
Downstream
Pum1
Celf3/4/6
Pum1
Dazap1
Pum1
Khdrbs2/3
Pum1
Celf3/4/6
Cnot4
Srsf1/9
FDR
C
miR-17/20/106 seed
Argonaute
miR-17/20/106 seed
Pumilio
m9U
m1A
UBNX2A
AGO2
PUM2

2.4 Discussion

We have presented an integrative model for predicting miRNA binding sites by training on sequencing assays that map biochemical interactions via AGO cross-linking and miRNA-mRNA ligation. We demonstrated that chimiRic can detect non-canonical miRNA-mRNA binding modes and significantly outperforms MIRZA for predicting the interacting miRNA for both canonical and non-canonical mRNA target sites. Moreover, chimiRic outperforms TargetScan, a leading target prediction method, for discriminating canonical seed sites that are bound by AGO from unbound sites. The feature representation of our AGO binding model exploits recent 3'-end sequencing data that identifies alternative 3'UTR isoforms and enables analysis of mRNA sequence signals in the vicinity of the miRNA binding sites, suggesting that other RBPs may collaborate with AGO to mediate miRNA-mRNA interactions.

ChimiRic directly predicts miRNA targeting by learning from miRNA binding data, whereas most existing algorithms infer miRNA targets and model their efficiency using mRNA expression changes in miRNA overexpression experiments in cell culture [49, 50]. One major issue with methods trained solely on gene expression changes is that the direct effects of miRNA regulation are confounded with secondary effects, leading to label noise in the learning problem. Since the true binding sites that mediate direct regulation are unknown in this setting, inference of miRNA targets involves "bootstrapping" from an initial set of assumptions of what constitutes a viable target. Furthermore, miRNA transfections in cell culture represent a non-physiological context for miRNA activity and may not accurately reflect endogenous targeting rules. Finally, miRNA binding can inhibit translational

efficiency of target mRNAs in addition to or instead of reducing mRNA abundance [69]. While previous global studies suggest that miRNA-mediated changes at the mRNA and protein levels are correlated [19, 70], these data also depend on miRNA overexpression in cell lines. For all these reasons, it is possible that what we have already exhausted what can be learned indirectly from mRNA expression changes due to miRNA perturbations—and from miRNA overexpression experiments in particular—and that new AGO CLIP and CLASH technologies for mapping direct interactions are required to advance our understanding of miRNA targeting in cells.

However, recent assays for mapping AGO sites and miRNA-mRNA interactions are technically difficult and present significant challenges for computational analysis and training of predictive models. CLASH and similar protocols that use RNA ligation to capture miRNA-mRNA interactions currently have very low ligation efficiency (only ~2% of reads are chimeric) [30, 31], suggesting that a large number of miRNA-mRNA interactions remain uncaptured. Some non-canonical interactions recovered by CLASH may be due to artifacts or biases in the ligation experiments, and one previous study found that incorporating chimeric reads into MIRZA did not significantly improve prediction performance [53]. Even in the more mature CLIP assays, data reproducibility is still limited and strongly affected by technical differences between various protocols (e.g. PAR-CLIP, HITS-CLIP, iCLIP) that produce protocol-specific biases [39] and by the potential false positives resulted from background binding [41]. In our experiments, we only trained on data sets with multiple biological replicates in order to ensure saturating coverage and to correctly label the mRNA sites as positive or negative. We further used a multi-task strategy to absorb dataset-specific differences into task-specific

models and learn a common model that captures general sequence signals and positional preferences of AGO binding. Although the extent of miRNA target context-specificity remains unclear [71, 72], it is still possible that there are true biological differences in AGO occupancy between cell types. Indeed, even directed perturbation of a single miRNA-mRNA interaction can lead to distinct changes in functional responses in different immune cell types [73]. Ultimately, as CLIP-based technologies mature and larger data sets accrue, the algorithmic approaches we present here may reveal the RNA sequence elements and trans-acting factors that mediate cell-type specific miRNA-mRNA interactions.

CHAPTER 3

THE EFFECT OF CELLULAR CONTEXT ON MIR-155 MEDIATED GENE REGULATION IN FOUR MAJOR IMMUNE CELL TYPES

Portions of this chapter are soon to be published and were written in collaboration with Jing-Ping Hsin, Gabriel Loeb, Christina Leslie and Alexander Rudensky[1].

3.1 Introduction

Cell type-specific regulation of gene expression, which is frequently mediated by commonly expressed sequence-specific transcription factors, is one of the foundational principles in developmental biology. Like transcriptional regulators, miRNAs with a proven, non-redundant role in cellular differentiation or function and their mRNA targets can be found in multiple cell types. In the immune system, a prime example of such miRNA is miR-155, whose expression is observed in functionally distinct T cell subsets, B cells, NK cells, macrophages, and dendritic cells, where it is induced in an activation or a differentiation stage-specific manner [74, 75]. miR-155 is also highly expressed in myeloid and lymphoid malignancies, where it plays an oncogenic role [76, 77]. Our recent study showed that miR-155 mediated regulation of an inducible target gene, Socs1, has widely differing cell type- and biological context-dependent functional significance in distinct types of lymphocytes [73]. Previously, we employed CLIP technology to identify

[1] As per the Cornell dissertation guidelines, the dissertation can include material that has been previously published or is soon to be published.

miR-155 targets through analyses of miR-155-suffienct and -deficient activated CD4 T cells [29]. Upon closer examination of the resulting datasets, we came across of a subset of conserved canonical miR-155 sites in expressed mRNAs that were not bound or regulated by miR-155 [29]. This finding raised the possibility that these sites may enable differential regulation of these targets in developmentally related immune cell types with a shared developmental origin.

However, recent analyses of immortalized human cell lines of different tissue origin including hepatocellular carcinoma, cervical cancer, and embryonic kidney cell transfected with hematopoietic and neuronal miRNAs (miR-155 or miR-124, respectively) showed that the majority of computationally predicted target mRNAs are repressed in a cellular context independent manner; a minor subset of differential regulation of a minor subset of miRNA targets observed in these cells was largely due to an alternative 3'UTR isoform usage with only two target mRNAs potentially regulated in an ApA-independent manner in miR-155 transfected cells [71]. While these experiments relied on overexpression of ectopic miRNAs, gene array and 3'UTR-seq analyses of mRNA expression in six different organs from miR-22–deficient and –sufficient mice were consistent with these results [71]. It can be argued, however, that differential regulation of mRNA targets by an endogenously expressed miRNA is more likely to be encountered in differentiated cell types of common developmental origin in response to a challenge or a developmental cue. Indeed, both endogenous cellular miRNAs and miRNAs encoded by Kaposi's sarcoma-associated herpes virus were found to regulate the expression of a sizable fraction of targets in distinct B cell lymphoma cell lines in a context-dependent manner [72]. However, the contribution of alternative 3'UTR isoform usage to miRNA-mediated regulation

of gene expression was not considered in this study [72]. Thus, it remains unknown whether endogenously expressed miRNA are capable of regulation of commonly expressed target genes solely in a cell context-independent or also in a cell context-dependent manner.

We sought to address this question through computational and comparative genome-wide molecular analyses of RISC bound mRNAs, using individual nucleotide resolution CLIP (iCLIP) [36], their 3'UTR usage (PolyA-Seq) and miR-155-dependent repression (RNA-Seq) in four key immune cell types – activated macrophages, dendritic cells, B cells, and CD4 T cells – isolated from miR-155-sufficient and -deficient mice. The analyses of the resulting datasets revealed notable cellular context-dependent miR-155 targeting and regulation of gene expression. While ApA contributed to differential miR-155 binding to some transcripts, in a larger number of cases, identical 3'UTR isoforms were differentially regulated across cell types. These results suggest ApA-independent and cellular context-dependent miR-155-mediated post-transcriptional regulation of gene expression reminiscent of transcriptional regulation by sequence-specific transcription factors. Furthermore, our study provides comprehensive comparative maps of miR-155 regulatory RNA networks as well as global miRNA-mediated AGO binding and genome-wide 3'UTR usage in key activated immune cell types.

3.2 Methods

3.2.1 Computational processing of iCLIP data

When we processed the iCLIP sequencing data, we first de-multiplexed the libraries based on the barcodes at 5′ end of the reads, and then preprocessed the reads using the cutadapt [78] software to remove the adaptor and low-quality bases. The remaining reads were aligned to the mouse genome (mm9) using the BWA aligner [79]. Multiple reads aligned to identical coordinates with the same random 7-mer in the barcode were considered as PCR duplicates and were merged into a single read to adjust for potential duplication biases. We then ran our peak-calling algorithm *CLIPanalyze* (manuscript in preparation) on the combined read coverage from all samples. The algorithm identified peaks by convolving the read coverage signal with the second derivative of a Gaussian filter. The locations where the convolved signal crosses zero correspond to the rising and falling edges in the original signal and these are used as boundaries for the peaks. Each peak was annotated with the corresponding gene name and its location within the gene (i.e. intron, CDS, 5′UTR, 3′UTR). Peaks within intergenic regions further than 5 kb downstream and 1 kb upstream from annotated genes were excluded from subsequent analysis. The peaks were then quantified by counting the number of uniquely aligned reads mapped within peak boundaries in each library. To filter the low-abundance peaks, we first restricted to peaks with supporting reads in at least 4 out of 8 samples in at least one cell type. For each individual cell type, we did a second round of filtering and only kept the peaks with total read counts within the top 10 percentile for the differential analysis. We then fit the read counts from those peaks using

negative binomial generalized linear models [80] with TMM normalization [81], and tested the significance of the difference in read counts between wild type and miR-155 KO samples with likelihood ratio test.

3.2.2 Computational processing of gene expression data

We preprocessed the paired-end reads using cutadapt to remove the adaptors and low-quality bases. The processed reads were then aligned to the mouse genome (mm9) using the STAR aligner [82]. To account for the variation in 3'UTR usage, we only counted the reads aligned to CDS for coding genes. The read counts per gene were further normalized as fragments per kilobase per million (FPKM) to represent the mRNA abundance.

Differential gene expression analysis was performed for microarray datasets of miR-142a-sufficient and -deficient B cells (GSE61919) [83] and bone marrow derived miR-142a-sufficient and -deficient dendritic cells (GSE42325) [84] using *limma* [85]. To estimate gene regulation mediated by miR-27a, differential gene expression analysis was performed on a RNA-Seq dataset of wild type and miR-27a-overexpressed CD4 T cells (GSE75909) [86].

3.2.3 Computational processing of PolyA-Seq data

The preprocessing, alignment, peak calling and quantification steps for the PolyA-Seq libraries were performed in the same way as the iCLIP libraries. Internally primed peaks were removed in the same approach as previously described [58]. The read counts were then fitted using the DEXSeq model [87] in

order to identify the differential usage of 3'UTR isoforms between conditions.

3.3 Results

3.3.1 Differential AGO2 iCLIP reveals context specificity of miR-155 targeting in activated immune cells

To comprehensively characterize the miR-155 regulatory network, we used iCLIP [36] to precisely map the miR-155 target sites, RNA-Seq to measure the repression levels of target genes, and PolyA-Seq [47] to map and quantify 3' UTR isoforms in B cells, dendritic cells, macrophages and CD4 T cells extracted from both wild type and miR-155 KO mice (Figure 3.1a). As previously reported [88–91], miR-155 expression was significantly increased upon immune activation in all four cell types, with peak induction levels observed at 24 h and extending to 48 h. We used Argonaute 2 antibody to immunoprecipitate RISC-bound RNA from cells activated for 48 h and generated iCLIP libraries from the isolated RNA captured both the microRNAs and their mRNA target sequences. Cellular abundances of mature microRNAs were estimated from reads aligned to the corresponding loci in primary microRNA sequences, which confirmed that miR-155 was the only major microRNA with significant change in expression between WT and miR-155 KO cells. By applying our CLIP processing pipeline *CLIPanalyze* (manuscript in preparation) to the genomic alignments after removal of potential PCR duplicates, we first identified peak regions in the combined read coverage track (WT and KO replicates) from all cell types. We then modeled the read counts within peaks using

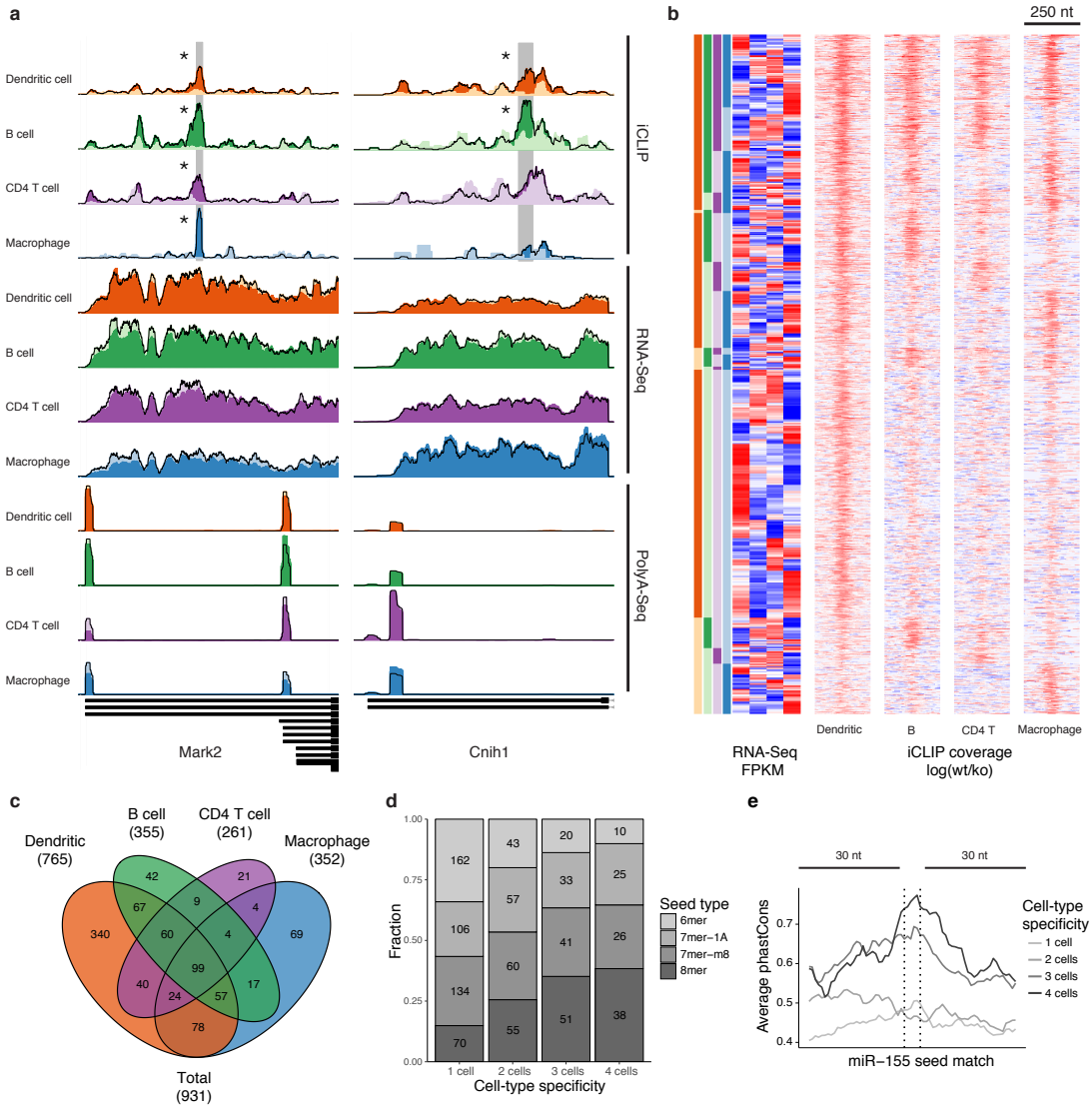
negative binomial generalized linear models [80] with TMM normalization [81] and determined the miR-155 dependent sites as peaks with (1) overlapping transcript annotation from RefSeq; (2) complementary match to miR-155 2-7 6-mer seed sequence; and (3) significantly higher read counts in wild type samples than miR-155 knockout samples (Benjamini-Hochberg adjusted $p < 0.025$). In total, 1,200 such sites were found in 999 genes across four cell types, including 796 (66.3 %) in 3′ UTRs, 386 (32.2 %) in CDS (coding sequence), and 18 (1.5 %) in 5′ UTRs. In particular, among these initial 1,200 sites, only 111 (9.25 %) were found to be miR-155 dependent in all four cell types, while the rest of the targets exhibited varying degrees of context specificity.

3.3.2 Differences in target mRNA and miR-155 abundance do not account for all miR-155 targeting specificity

One obvious explanation for the observed context specificity could have been that some of the cell-type specific miR-155 target genes were not expressed or were expressed at very low levels in the other cell types. Indeed, when a gene contains a miR-155 target specific to one cell type, its mRNA expression in that cell type also tended to be higher than in those where the target did not show differential iCLIP signal. When we restricted the comparison between cell types to the co-expressed genes (RNA-Seq FPKM > 1 in all cells and < 16-fold difference between any two cell types), 931 target sites in 778 co-expressed genes remained and most of the context specificity was preserved (Figure 3.1b and 3.1c). Therefore, the base mRNA expression differences alone cannot fully account for the observed cell context-specific targeting.

The difference in miR-155 abundance after immune stimulation across cell types can also partially explain the cell context specificity of miR-155 targeting – the largest number of cell-type specific target sites were found in dendritic cells, where miR-155 expression was also the highest (Figure 3.1b). The number of miR-155 dependent sites identified in each cell type is consistent with relative miR-155 expression (Figure 3.1c), suggesting that some context-specific sites may have weaker affinity to miR-155 and, therefore, can only be regulated in the presence of higher miR-155 levels or other cellular factors. Indeed, when we categorized the miR-155 targets by the number of cell types that they are present in, the proportion of sites with only 6-mer complementarity was significantly lower for target sites present in more cell types than those present in fewer cell types (Fisher's exact test $p < 2.57e\text{-}10$), and the proportion of sites with 8mer complementarity significantly higher (Fisher's exact test $p < 1.79e\text{-}9$; Figure 3.1d). Similar to previous observations [72], the sequences surrounding shared sites also showed significantly higher evolutionary conservation than the sequences around cell-type specific sites (Figure 3.1e). Nevertheless, large numbers of context-specific targets are still present in cell types with lower miR-155 expression, suggesting that other cellular factors or potentially alternative cleavage and polyadenylation (ApA) play a role in cell-type specific targeting.

Figure 3.1: **miR-155 mediated Argonaute binding occurs at distinct sites in four immune cell types.** (A) Examples of universally bound and differentially bound miR-155 sites across all 4 cell types. Normalized read coverage tracks of iCLIP, RNA-Seq and PolyA-Seq libraries are shown for each cell type, in which dark and light colors correspond to the wild type (WT) and miR-155 knockout (KO) samples. iCLIP peaks are defined by the grey shade in the background, while asterisks designate the cell types with significant (FDR < 2.5%) difference between WT and KO coverage. (B) Summary of miR-155 dependent sites in co-expressed genes, including 3'UTR, CDS, and 5'UTR sites, identified by differential iCLIP. Each row in the heatmap represents 250 bp around a miR-155 6-mer seed match, whereas the colors represent the log ratios between normalized WT and KO iCLIP coverage per base. Heatmap for RNA expression (WT RNA-Seq log10 FPKM, normalized by row) of the same genes containing the miR-155 sites is shown side-by-side. Sites are categorized according to their binding specificity across 4 cell types, while the order within each category are determined by hierarchical clustering of RNA-Seq FPKM values for corresponding genes. (C) Venn diagram of miR-155 dependent iCLIP sites in co-expressed genes. (D) Seed type composition of miR-155 dependent sites in co-expressed genes. (E) Average base-wise phastCons scores (for multiple genome alignments between mouse and other 39 placental mammals) of miR-155 dependent sites in co-expressed genes.


a
Dendritic cell
B cell
CD4 T cell
Macrophage
iCLIP
RNA-Seq
PolyA-Seq
Mark2
Cnih1
b
250 nt
Dendritic
B
CD4 T
Macrophage
RNA-Seq FPKM
iCLIP coverage log(wt/ko)
c
Dendritic (765)
B cell (355)
CD4 T cell (261)
Macrophage (352)
Total (931)
d
Fraction
Cell-type specificity
1 cell
2 cells
3 cells
4 cells
Seed type
6mer
7mer−1A
7mer−m8
8mer
e
30 nt
Average phastCons
miR−155 seed match
Cell-type specificity

3.3.3 miR-155 targeting is unlikely to be influenced by endogenous RNA competition

The "competitive endogenous RNA (ceRNA)" hypothesis [92] proposes that transcripts with common microRNA target sites compete with each other for regulation, which may explain the biological function of some long non-coding RNAs. There has been growing experimental evidence that certain long non-coding RNAs [93] and circular RNAs [94, 95] contain large numbers of microRNA target sites and may function as microRNA "sponges", particularly in neurons. However, when we examined miR-155 target sites in mRNA along with ones within intronic regions and non-coding RNAs, we found the vast majority of coding and non-coding RNAs only contained one or two miR-155 target sites in all four cell types, with the maximum of six sites found only in single gene, Picalm. As circular RNAs are generally formed by back-splicing of consecutive exons [96], we therefore find little evidence of circular RNAs that "sponge" miR-155 in these four immune cell types.

We also attempted to estimate the fraction of miR-155/AGO complex bound by a given transcript in each cell. Assuming that iCLIP counts are a reasonable proxy for miR-155/AGO binding, we estimate that the most bound transcript in a given cell binds ~3-10% of the transcript bound complex. This suggests that these rare already highly expressed transcripts would need to be dramatically up-regulated to significantly affect overall miR-155 binding within the cell. Interestingly, the most bound targets are different for each cell type, even between these closely related immune cells. Among the predominant target miR-155 genes in dendritic cells was Cd274, encoding the inhibitory receptor ligand PD-L1, and in macrophages Msr1, encoding macrophage scavenger

receptor 1.

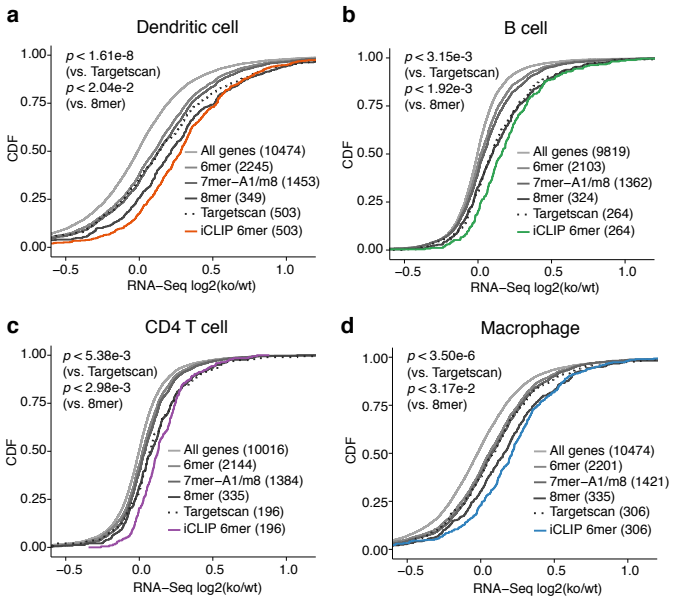


Figure 3.2: **miR-155 represses distinct sets of genes in four immune cell types.** In dendritic cells (A), B cells (B), CD4 T cells (C) and macrophages (D), the distribution of gene-level RNA-Seq expression changes between miR-155 KO and WT cells is shown in the form of cumulative distribution functions (CDFs) in different sets of genes. Gene sets include all expressed genes, genes with 3′ UTR miR-155 6-mer / 7mer-A1 / 7mer-m8 / 8-mer seed matches and genes containing 3′ UTR miR-155 dependent iCLIP sites with 6-mer seed matches (FDR < 2.5%). Predicted miR-155 target genes with top context++ scores from Targetscan 7.0 (same number as the miR-155 target genes identified by differential iCLIP) are also shown.

3.3.4 miR-155 mediated gene regulation is consistent with the context specificity of iCLIP-defined targets

Next, we analyzed the extent of regulation induced by miR-155 dependent targets identified by differential iCLIP. We used mRNA expression changes between wild type and miR-155 knockout cells to estimate the extent of miR-155

regulation per gene. Consistent with previous studies [97], the significance of miR-155 dependent iCLIP sites in 3′ UTRs correlated with the extent of regulation of corresponding genes, which was not the case for CDS sites. Therefore, for further analyses of the effect of miR-155 on gene regulation we only considered miR-155 targets in 3′ UTRs.

In all four immune cell types, we first examined the distribution of mRNA expression changes of potential target genes defined by miR-155 seed matches in the 3′ UTRs. Consistent with well-known microRNA targeting principles, the extent of miR-155 regulation increased with higher 3′UTR seed complementarity, from 6-mer to 7-mer-A1/m8 to 8-mer [26]. Still, genes with miR-155 dependent iCLIP sites in the 3′UTRs displayed significantly stronger regulation even when compared to the most potent predicted target genes with 8-mer seed matches in 3′ UTRs (Figure 3.2). We also compared the iCLIP-defined target genes to same number of genes containing sites with top *context*++ scores from Targetscan 7.0 [49]. While the extent of regulation in the top ~10% of the distribution was similar for both sets of genes, the iCLIP-defined target genes overall show significantly stronger regulation compared to Targetscan predictions (Figure 3.2). These results again suggest that miR-155 mediated gene regulation across different cellular contexts is more accurately captured by differential iCLIP assays than cell-type agnostic sequence-based predictions.

We have previously reported that up to 40% of miR-155 targets identified by differential AGO2 HITS-CLIP in CD4 T cells are non-canonical [29], i.e. without complementary match to miR-155 6-mer seed. More recent studies by other groups using CLIP-based assays with RNA ligation [30, 31, 43, 44] to recover

miRNA-target interactions have also suggested widespread non-canonical targeting with 3' end complementarity. In line with previous reports, we found non-canonical sites consisted of about 25%-45% of identified AGO2-bound miR-155 sites in the four immune cell types. The majority of non-canonical sites were bound in only one cell type, which was consistent with the observation of canonical seed type composition. Similarly, when we compared the average iCLIP read coverage around the canonical and non-canonical miR-155 sites, we found that the difference between wild-type and miR-155-deficient libraries was much smaller in non-canonical sites, suggesting that the non-canonical sites have weaker affinity to RISC binding. We found multiple genes significantly repressed by miR-155 with only non-canonical target sites in 3' UTR, albeit the regulation of non-canonical targets was significantly weaker overall than canonical targets even with the most stringent FDR cutoff.

To further dissect the cell-context specificity of miR-155 regulation, we performed pairwise comparisons across the four immune cell types to assess the extent of regulation of common and cell-type specific miR-155 targets (Figure 3.3). In each immune cell type, miR-155 target genes identified by differential iCLIP always displayed significantly stronger regulation than those specific to other cell types, with a few exceptions involving B and CD4 T cells, where fewer cell-type specific targets and generally weaker regulation were observed. Notably, cell-type specific target genes displayed significantly less pronounced regulation compared to common target genes, consistent with the weaker seed complementarity and lower sequence conservation associated with cell-type specific target sites (Figure 3.1d and 3.1e).

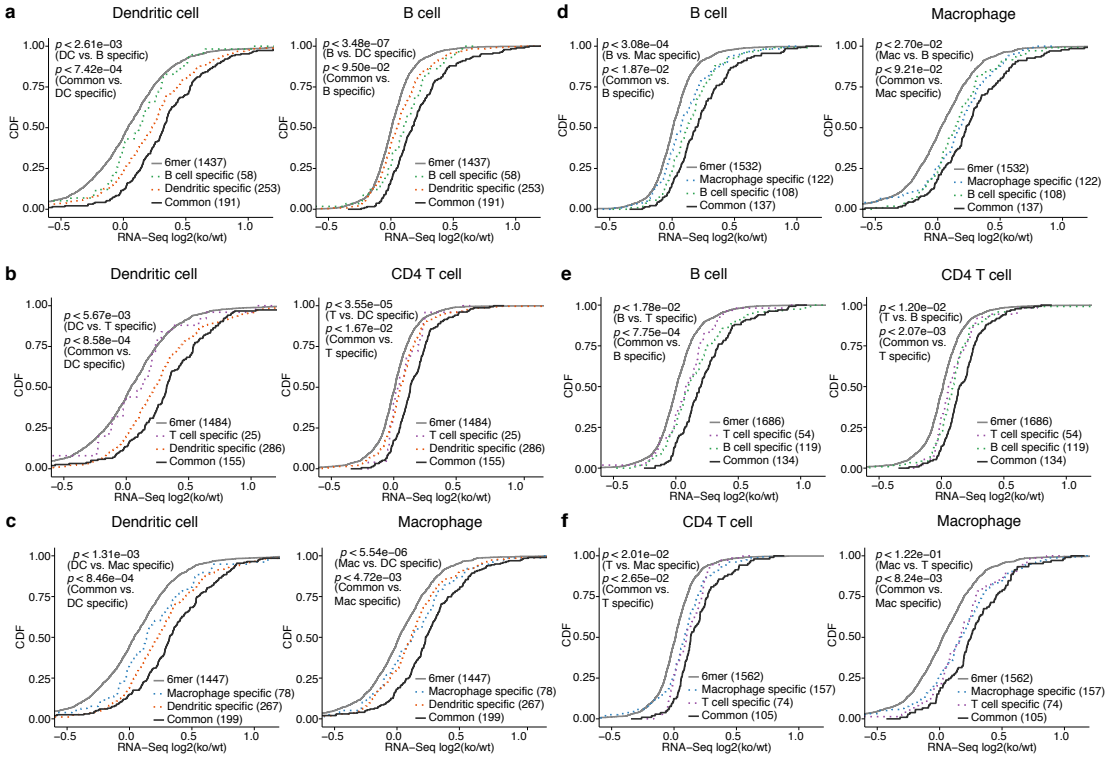


Figure 3.3: **Context-specific miR-155 targeting leads to differences in gene regulation between cell types.** For all six pairwise comparisons across four immune cells, de-repression of genes containing common (solid lines) and cell-type specific (dotted lines) 3′ UTR miR-155 dependent iCLIP sites is shown in the form of CDFs. Genes with 3′ UTR miR-155 seed matches are also shown as reference. Only co-expressed genes (WT RNA-Seq FPKM > 1 and difference < 16 fold) are included in each pairwise comparison. In each plot, two p-values from one-sided KS tests are shown. First one corresponds to the comparison between all miR-155 target genes identified in this cell type and genes only targeted in the other cell, while the second one corresponds to the comparison between the common target genes and target genes specific to this cell type.

3.3.5 Alternative polyadenylation has limited contribution to cell-type specific miR-155 targeting

Another potential explanation for the observed cell type-dependent regulation of gene expression by miR-155 is alternative polyadenylation. Previous studies [98, 99] have shown that multi-UTR genes increase the usage of shorter isoforms

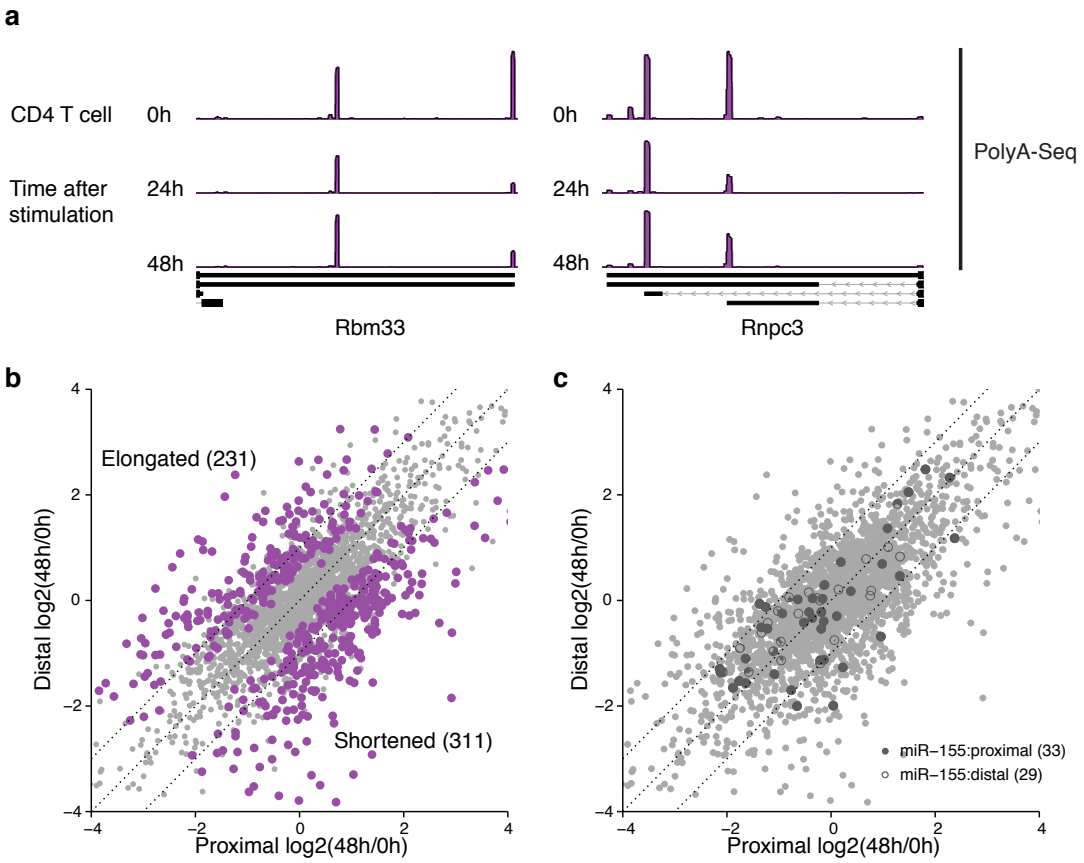


Figure 3.4: **PolyA-Seq captures change in 3′ UTR isoform usage during CD4 T cell activation.** (A) Two examples of 3′ UTRs with significant (FDR < 5%) isoform usage changes during CD4 T cells activation. Tracks represent normalized PolyA-Seq read coverage at 0h, 24h and 48h after activation. (B) The changes in 3′ UTR isoform usage for 3′ UTRs with two major isoforms at 48 h after CD4 T cell activation. The ones undergoing significant usage changes were highlighted. (C) Same as (B), but highlighting the two-isoform 3′UTRs containing target sites of miR-155. The 3′UTRs containing proximal (solid shapes) and distal (hollow shapes) miR-155 target sites were marked separately, as well as the corresponding numbers.

in activated immune cells, specifically T lymphocytes, simultaneously with the increase in miR-155 expression, which has been suggested by some as a potential mechanism to evade miRNA-mediated regulation. We performed PolyA-Seq in naïve CD4 T cells as well as their activated counterparts after *in vitro* stimulation with CD3 and CD28 antibodies for 24h and 48h (Figure

3.4a). Although differential analysis [87] indeed revealed widespread changes in 3'UTR isoform usage with a significant shift towards shorter isoforms in activated cells both at 48h (Figure 3.4b), markedly increased usage of longer isoforms upon activation was also observed for a sizable group of transcripts (~40%). A focused analysis of the two-isoform 3'UTRs targeted by miR-155 did not suggest preferential shortening of transcripts that contained a miR-155 binding site in the long isoform (Figure 3.4c). Changes in 3'UTR length thus did not appear to significantly relieve miR-155 mediated targeting upon T cell activation.

To investigate whether alternative polyadenylation contributed to cell-type specific targeting we performed PolyA-Seq in all four immune cell types. The PolyA-Seq FPM was well correlated with RNA-Seq FPKM for single-UTR genes, suggesting that PolyA-Seq is capable of quantifying 3'UTR isoform expression levels. Differential analysis [87] in all four cell types showed that 2,703 out of 3,460 co-expressed multi-UTR genes displayed some extent of alternative polyadenylation (Figure 3.5a). miR-155 targets were significantly enriched in differentially used multi-UTR genes compared to the other genes (Fisher's exact test $p <$ 2.2e-16, Figure 3.5b). Since PolyA-Seq libraries were generated for both wild type and miR-155 KO cells, the data also allowed us to assess miR-155 regulation at the level of 3'UTR isoforms. In agreement with previous observations [26], regulation of a 3'UTR isoform by a given miR-155 target site negatively correlated with its distance from the 3'UTR end, suggesting the potential of ApA as a mechanism for context-specific miR-155 regulation. Indeed, in multi-isoform 3'UTRs, we observed that the extent of gene-level miR-155 regulation generally increases with higher usage of ApA isoforms containing miR-155 target sites in individual cell types as previously reported

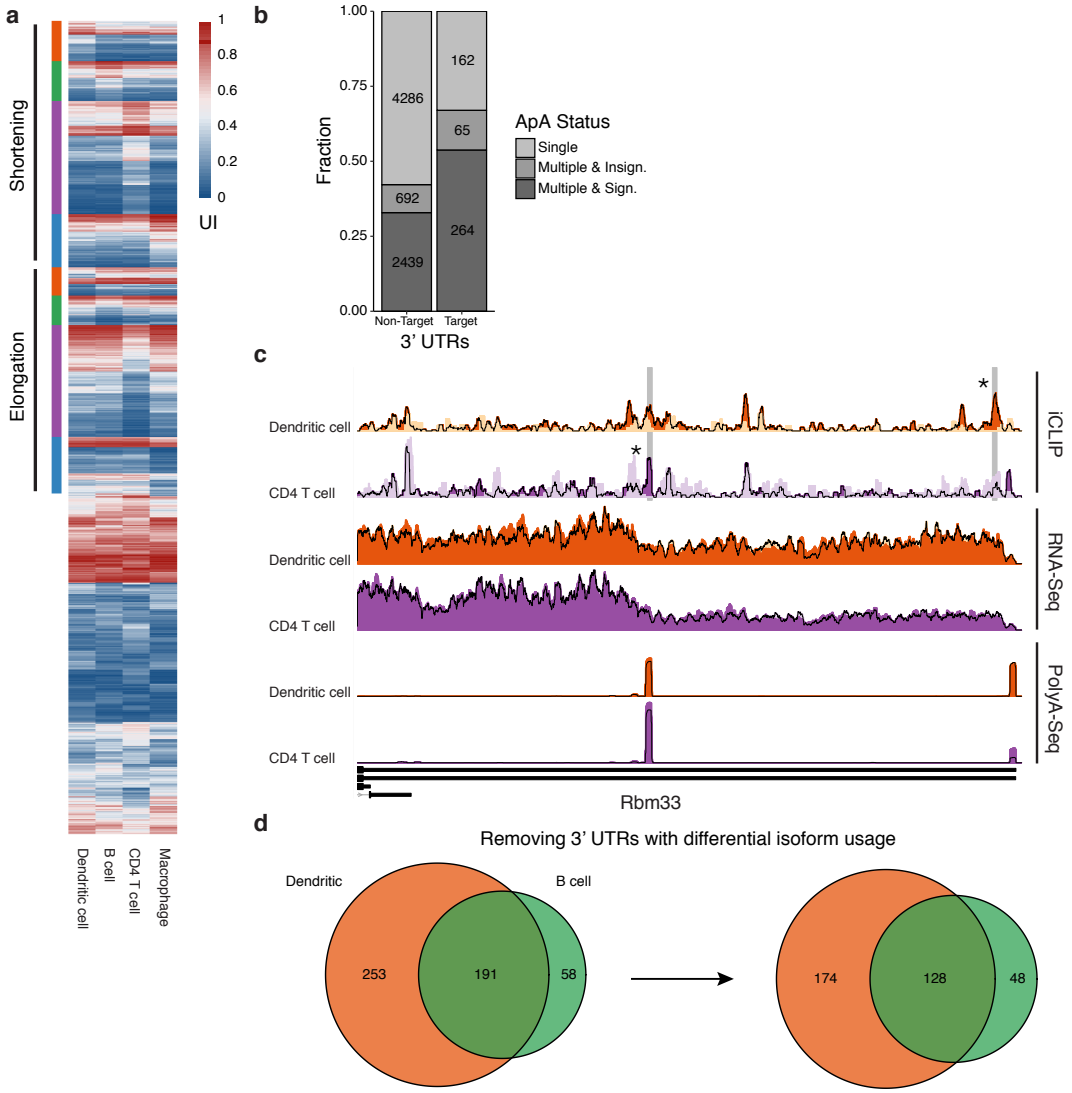


Figure 3.5: **The role of alternative polyadenylation in cellular context dependent regulation of gene expression by miR-155.** (A) A heatmap showing the usage changes in multi-isoform 3'UTRs across all four cell-types. The usage index (UI) represents the usage of the shorter isoform for two-isoform 3'UTRs, while for 3'UTRs with more isoforms it corresponds to the usage of the one shorter isoform with the most significant usage change. (B) Composition of 3'UTRs with single isoform, multiple isoforms with and without context-specific usage, divided into two categories depending on miR-155 targeting. (C) iCLIP, RNA-Seq and PolyA-Seq read coverage tracks in Rbm33 3'UTR, an example of the co-occurrence of differential ApA and context-specific miR-155 targeting between dendritic cell and CD4 T cell. (D) Venn diagram shows the shared and context-specific 3'UTR miR-155 target genes between dendritic cell and B cell, before and after removing genes with differential ApA usage in multi-isoform 3'UTRs.

[71]. We also observed examples of co-occurrence of ApA and context-specific miR-155 binding through pairwise comparison between cell types (Figure 3.5c). However, in most cases, the change in isoform usage between cell types was less than 10%, while overall expression changes of miR-155 and target mRNAs had a much larger dynamic range. Therefore, the majority of the observed context-specific targeting cannot be attributed to alternative polyadenylation (Figure 3.5d).

3.3.6 Ago iCLIP characterizes functional target sites of other miRNAs

Our Ago iCLIP data also allowed characterization of target sites for other miRNAs expressed in the four immune cells. The latter relied on computational seed sequence analysis within iCLIP peaks in the absence of a genetic control, i.e. iCLIP and RNA-seq analysis of corresponding miRNA-deficient cells. When we ranked iCLIP peaks containing miR-155 6-mer seed matches by the normalized read counts in wild-type libraries, ~75%-95% in the top 10% of peaks overlapped with miR-155 dependent sites defined by differential iCLIP. We therefore reasoned that stringent read count cutoffs could yield reliable sets of targets for miRNAs other than miR-155. Using the wild-type libraries, we defined the top target sites for miR-142a-3p and miR-27a-3p, which both play key regulatory roles in immunity [83, 84, 86] and were highly expressed in the four immune cells. When we used publicly available gene expression data with perturbed miR-142a [83, 84] and miR-27a [86] expression in mouse immune cells, we found that similar to miR-155, the target genes defined by 3'UTR

iCLIP sites with top read counts in wild-type libraries showed significantly stronger repression than cell-type agnostic sequence-based predictions (Figure 3.6, one-sided KS test), which suggests that they indeed defined an accurate set of top miRNA targets in the respective cellular context.

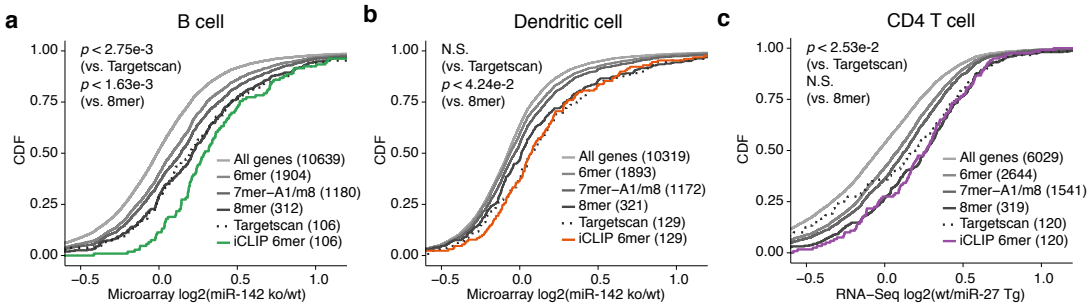


Figure 3.6: **Top iCLIP target sites of other miRNAs induce significant gene repression.** mRNA expression changes in B cells (A) and dendritic cells (B) with miR-142a KO and in CD4 T cells with miR-27a overexpression (C) are shown as CDFs for different gene sets. Gene sets consist of all expressed genes, genes with 3'UTR seed matches (6mer, 7mer-A1, 7mer-m8, and 8mer), and genes containing 3'UTR iCLIP sites with 6mer seed matches and most reads in wild-type libraries. Predicted miRNA target genes with top *context*++ scores from Targetscan 7.0 (same number as the target genes defined by wild-type iCLIP) are also shown.

CHAPTER 4

DISCUSSIONS

4.1 Computational miRNA target prediction

Computational miRNA target prediction methods share the same issue that it is relatively easy to predict miRNA target sites with the highest affinity, but the accuracy of prediction rapidly drops when it comes to weaker targets. A recent study has also suggested that a large proportion of predictions made by miRNA target prediction methods are likely false positives [100]. As shown by our study on miRNA context specificity, the variations in miRNA abundance have a significant impact on the regulation of target sites with weaker affinity to miRNAs. Therefore, one potential solution to this issue is taking the miRNA expression levels into account when predicting the target sites in a specific cell type.

Our original chimiRic model was partially restricted by the limited number of high-quality AGO CLIP and CLASH data sets available. Since the publication, more ligation-based data sets have been generated [43, 44], and we have now recognized the importance of control libraries in correcting the intrinsic biases of CLIP data [37, 41]. Besides better data quality, miRNA target prediction algorithms may also benefit from novel machine learning methods. Algorithms based on deep neural network have proved to be powerful tools for modeling genomic data [101, 102], and recently they have also been applied to predicting miRNA target sites based on AGO CLIP data [103]. In addition, more powerful machine learning algorithms can potentially help us revisit the rich miRNA perturbation expression data sets accumulated over the years [104]

and recover features of miRNA target sites that were previously neglected in individual studies.

4.2 Context specificity of miRNA regulation

In our study on miR-155 regulation in immune cells, we observed significant context specificity of miRNA regulation, contradicting the results of previous perturbation assays in cell lines and in whole tissues [71]. A possible explanation is that the cell lines used in previous study are not developmentally related, and that bulk gene expression in whole tissues may limit the discovery of more subtle differences in gene regulation. We only observed a small number of context-specific miR-155 targets potentially affected by ApA, although in theory ApA can play a more significant role in tissues where ApA events are more prevalent, like in brain [58].

Limited by the relatively small number of context-specific miR-155 targets in our data, we did not fully explore other ApA-independent mechanisms of miRNA context specificity. We hypothesize that regulation by RBPs differentially expressed between cell types can play a significant role. More than 1,500 RBPs have been identified in human [105], and individual studies have identified multiple RBPs as either enhancers [64, 106] or inhibitors [63] of miRNA regulation. There has been attempts at systematic characterization of RBPs' impact on miRNA regulation [107], but they were largely limited by the lack of direct measurement of RBP binding. A large number of high-throughput data sets of RBP binding sites have been generated in recent years, which may help resolve this issue. In particular, the ENCODE eCLIP experiments

have characterized the binding regions of 126 RBPs in human cells [108]. Comprehensive analysis of RBP co-binding profiles would greatly enhance our understanding of how RBPs modulate miRNA regulation and how they can contribute to cell-type specific miRNA targeting.

mRNA modifications, such as adenosine methylation (N^6-methyladenosine, m^6A), can be another potential factor contributing to context-specific miRNA regulation. It has been previously observed that >70% of m^6A residues are present in the last exons of transcripts [109], which significantly overlap with miRNA target sites. A different modification (N^6,2′-O-dimethyladenosine, m^6A_m) at the first encoded nucleotide adjacent to the 5′ cap has also been suggested to stabilize mRNAs by preventing miRNA-mediated degradation [110]. Comprehensive mapping of transcriptome-wide RNA modifications can help us better understand the various mechanisms involved in miRNA regulation.

4.3 Technical advances in the detection of RBP binding sites

Continuous efforts have been made in order to improve the efficiency of CLIP protocols. The "on beads" PAR-CLIP protocol performs all RNA adapter ligation steps while RNA fragment is still cross-linked to the RBP, reducing the library preparation time by three days compared to the original PAR-CLIP workflow [111]. The infrared-CLIP (irCLIP) protocol [112] uses an infrared-dye-conjugated and biotinylated RNA ligation adapter to eliminate the need for radioisotopes to visualize protein-RNA complexes. In combination with improved RNA digestion, purification and reverse

transcription procedures, this protocol greatly increases the efficiency of CLIP library preparation and enables direction of protein-RNA interactions using far fewer cells.

One intrinsic limitation of CLIP protocols is that the stringency of washing to remove non-specific protein interactions is restricted by the strength of protein-antibody interactions. Therefore, an additional protein gel purification step is necessary, which further reduces the efficiency of CLIP protocol [45]. Moreover, specific antibodies may not be available for certain RBPs. Instead of relying on immunoprecipitation, an alternative strategy is to use CRISPR to insert protein purification tags (such as HIS, Bio/BirA, or TAP-TAG) next to the genomic loci of RBPs, forming fusion proteins that can be captured by stronger covalent interactions and allow harsher washing to remove non-specific interactions. Notably, the recently developed Halo Tag has highly specific affinity to the HaloLink resin for protein capture [113], therefore offering great potential for further improvement of protein capture specificity. For these assays, preliminary experiments would be necessary to ensure that the added tags do not disrupt the *in vivo* binding affinity and biological functions of the original RBPs.

Furthermore, researchers are also exploring methods that does not involve protein pulldown and crosslinking. One such method, TRIBE, fuses the deaminase domain of a RNA-editing enzyme, ADAR, to the RBP of interest [114, 115]. The fusion protein introduces A-to-I RNA editing near RBP binding sites, which can be detected by RNA sequencing. Since RNA-Seq requires much fewer cells than CLIP, TRIBE enables characterization of differences in RBP regulation between small populations of cell subtypes. On the other hand, the

potential bias in the selection and efficiency of editing sites needs to be carefully evaluated, since current results suggest that the ADAR deaminase domain may retain some of binding preference of the original enzyme [114, 115].

4.4 Unsolved questions

Multiple important questions regarding the physiological consequences of miRNA regulation remain unsolved. It is still unclear how miRNAs can have massive phenotypes when the majority of miRNA targets are only mildly repressed [19, 70]. Theoretical and experimental analyses have suggested that miRNA regulation can reduce the variations in mRNA and protein abundances [116], which may play more important regulatory roles than simple repression of target gene expression. The competitive endogenous RNA (ceRNA) hypothesis [92] proposes that transcripts with common miRNA target sites compete with each other for regulation, which provides another attractive theory for the mechanism of miRNA regulation. It has been found that certain long non-coding RNAs [93] and circular RNAs [94, 95] contain large numbers of miRNA target sites and may be the miRNA "sponges" that the ceRNA hypothesis proposes. On the other hand, quantitative modeling and measurements showed that the majority of active miRNAs are probably not susceptible to ceRNA competition [42, 117]. To date, the ceRNA hypothesis remains controversial because of these conflicting observations [118]. Overall, 25 years after its discovery, miRNA regulation continues to provide challenges and opportunities to both experimental and computational biologists.

APPENDIX A

THE CLIPANALYZE DATA PROCESSING PIPELINE

Portions of this chapter first appeared in Park et al. [119] and were written in collaboration with Sun-Mi Park, Christina Leslie and Michael Kharas [1].

A.1 Introduction

Investigations in RBP biology are increasingly dependent on CLIP and related high-throughput sequencing protocols. On the other hand, due to various biases and noises present in CLIP libraries, careful computational analysis is necessary for correct interpretation of CLIP data. Several CLIP data analysis pipelines have been developed before, but to date none of them have been widely adapted by the research community, since most of them were limited to certain CLIP protocol variants. We have implemented a software pipeline that performs standard CLIP data processing procedures including peak identification, annotation and quantification in a highly efficient manner. In addition, our pipeline accounts for biases in CLIP libraries that come from two major sources, PCR duplication and non-specific protein-RNA interactions. We release our pipeline as a R package *CLIPanalyze*, and the source code is publicly available at `https://bitbucket.org/leslielab/clipanalyze`.

[1] As per the Cornell dissertation guidelines, the dissertation can include material that has been previously published or is soon to be published.

A.2 Methods

A.2.1 Pre-processing and alignment

To adjust the potential biases generated by uneven PCR amplification, it has become a common practice of various CLIP protocols to include a randomized barcode sequence in the adapter or reverse transcription primer [27, 36, 37]. Our pipeline stripped the random barcodes from read sequences and attached them to read names in FASTQ files. After read alignment, multiple reads mapped to identical coordinates with the same random barcode were considered as PCR duplicates and were merged into a single read to adjust for potential duplication biases. In case that the CLIP libraries were constructed without random barcodes [42, 120], our pipeline also supported only using the alignment coordinates to remove PCR duplicates.

A.2.2 Peak calling

Our peak calling approach was inspired by the edge detection algorithm in computer vision, where sharp changes in brightness in an image are detected as edges of an object by computing the rate of change of the intensity gradient. To identify peaks, we first combined the reads from all of the CLIP libraries together. We then constructed a 1D signal profile of read coverage, $K[x]$, which contains cumulative read counts for each position x from all CLIP libraries. To simultaneously smooth and identify edges in the signal, this profile was convolved with a kernel derived from the second derivative of a Gaussian (g''_D),

with a mean of 0, standard deviation of 1, and customized bandwidth:

$$g''_D[x] = \left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

The customized bandwidth parameter allows adaptive peak calling across different CLIP data sets, since the typical peak sizes in CLIP libraries can be highly variable depending on the RBPs and experimental conditions [37]. The edges in the original signal are located at the zero-crossings of the convolved signal:

$$(K \times g''_D)[x] = \sum_{n=-m/2}^{m/2} K[x+n] \times g''_D[n]$$

The zero-crossings of the second derivative that switch from positive to negative indicate the edges that start a peak, and the points that switch from negative to positive identify the ends of each peak.

A.2.3 Peak annotation and quantification

Each CLIP peak was annotated according to the RefSeq gene annotation. Genes with multiple transcripts were reduced to a unified gene model, which is the union of all annotated exons. To account for possible gene structure variations that were not annotated, for each gene we extended the first exon 1 kb upstream and the last exon 5 kb downstream. Using these models, each CLIP peaks was then annotated to a specific genomic region within the gene (CDS, intron, 5′ UTR, and 3′ UTR) that the peak overlapped with. If a genomic region can be assigned to multiple categories, we assigned it to one of them with a customized priority order. The default order is: 3′ UTR, CDS, 5′ UTR, and intron. Peaks that mapped to multiple genomic regions were assigned to the region with maximum overlap. Once the peaks were identified and annotated, we then

quantified peaks in each experiment by counting the number of reads from each experiment that overlapped with each peak. Reads that overlapped more than one peak were assigned to the peak with which it had maximum overlap.

A.2.4 Normalization against control libraries

Since non-specific binding events are prevalent in CLIP [37, 41], it is highly recommended to normalize the CLIP-identified peaks against control libraries. For each peak, we used negative binomial generalized linear model [80, 121] to fit the difference in read counts between CLIP and control libraries. This approach can be easily adapted to more sophisticated experimental designs with additional factors. Considering the fact that the CLIP libraries intrinsically have more reads in peaks than the control libraries, we estimated the library sizes using the number of reads outside of CLIP peaks per gene instead of using the number of reads within peaks. The p-values from differential tests can then be used to represent the confidence of each peak.

A.3 Example: Computational analysis of MSI2 CLIP-seq in K562 cells

Earlier versions of *CLIPanalyze* pipeline has been applied to CLIP-seq data sets generated by multiple studies [29, 119, 122]. Here we describe details of the computational analysis performed for MSI2 HITS-CLIP [119]:

The Musashi (MSI) family of RNA-binding proteins, including MSI1 and

MSI2, contribute to the control of symmetric and asymmetric stem cell division, regulate stem cell function, and play a role in cell fate determination [123]. MSI proteins are thought to function by binding to the 3′UTRs of target mRNAs at a consensus sequence and then blocking translation by hindering access of the poly-A–binding protein to the elongation initiation complex [124]. In particular, MSI2 is an important modulator of proliferation and differentiation in both normal HSCs and in myeloid malignancies. Although MSI2 is most highly expressed in the primitive hematopoietic compartment, and MSI2 overexpression drives quiescent HSCs out of G0 and into cycle [125], it remains unclear whether and how MSI2 affects HSC self-renewal and commitment under homeostatic conditions. Furthermore, the critical RNA-binding targets of MSI2 in hematopoietic cells that regulate self-renewal and lineage commitment remain to be uncovered.

In order to globally capture the direct RNA targets of MSI2 in hematopoietic cells, we performed HITS-CLIP in K562, a human chronic myeloid leukemia cell line. We overexpressed FLAG-tagged MSI2 protein in parallel to a control vector lacking the MSI2 cDNA in K562 cells, and HITS-CLIP libraries were generated using anti-FLAG M2 antibody. Since FLAG-tagged MSI2 protein was not expressed in control cells, HITS-CLIP reads from the control sample were generated by non-specific binding of the antibody and other sources of background noise. A fraction of reads in samples with overexpressed MSI2 also came from these noise sources [27]. Therefore, we were interested in identifying HITS-CLIP peaks with significantly higher read counts in MSI2-overexpressing cells relative to control, as they are likely to be the real MSI2-binding sites. However, as mRNA expression levels may change between different conditions, differential read counts at a site can be caused either by a change in transcript

abundance or by differential MSI2 binding. To identify real differential binding events, it is necessary to integrate gene expression data into the analysis. We jointly modeled read count data from HITS-CLIP and RNA-Seq with a generalized linear model. We represented the read count from a window containing peak i in sample j as K_{ij}. Here, the read count represents either HITS-CLIP reads or RNA-Seq reads in the window, depending on sample j. It is assumed that K_{ij} follows a negative binomial distribution, which has been widely used in modeling read count data [80].

For each peak i, the expected value of K_{ij} (denoted by μ_{ij}), is fit via a logarithmic link by the following model:

$$log\mu_{ij} = \beta_i^0 + \beta_i^{CLIP} X_j^{CLIP} + \beta_i^{OE} X_j^{OE} + \beta_i^{CLIP:OE} X_j^{CLIP:OE} + log\tilde{N}_j$$

Here $\tilde{N}_j$ represented scaled library size of sample j, which was the total read count in sample j scaled by the weighted trimmed mean of log expression ratios. It was included as a normalization factor. After normalization, the logarithm of this variable is decomposed into four factors, where the regression coefficients have the following interpretation: β_i^0 represents the baseline log expression level measured by the window at peak i; β_i^{CLIP} represents the baseline log read count ratio of CLIP reads to RNA-Seq reads at peak i; β_i^{OE} represents the effect of MSI2 overexpression on read counts caused by mRNA expression changes. Finally, the interaction term $\beta_i^{CLIP:OE}$ represents differential MSI2 binding caused by overexpression; this coefficient will be non-zero if there is differential binding even after controlling for differential mRNA expression. Factors X_j^{CLIP}, X_j^{OE} and $X_j^{CLIP:OE}$ equal 1 or 0, depending on the condition and library type of sample j. To test whether the interaction term is 0, we fit data to both the full model and a reduced model without the interaction term. Then

the deviances of two models were used to conduct a likelihood ratio test. If there is no true differential binding effect, the difference in deviances between the nested models should be small compared with a χ^2 distribution with one degree of freedom. In this way, we were able to characterize the significance of differential binding with a p-value. Finally, we defined MSI2-binding sites as sites with Benjamini-Hochberg– adjusted p-value < 0.1 and $\beta_i^{CLIP:OE} > 0$. The above analysis was conducted using the *edgeR* package [80].

We found 1,097 unique targets that have at least one significant MSI2-binding site (adjusted p-value < 0.1 with a corrected CLIP log fold change of two or more). We observed the binding was distributed between the coding sequence (CDS) and the UTRs (56% and 44%, respectively). We then queried the MSigDB signatures with GSEA [126] using the full list of CLIP targets ranked by fold change to understand the functional classification of MSI2's targets. MSI2 binding was positively enriched for 668 gene sets (FDR < 0.01). We then examined the genes sets and categorized them into two modules, "RNA regulation and electron transport" and "Signaling and development" based on the overlaps between gene sets. For instance, within the "RNA regulation and electron transport module", genes sets containing genes that are normally down-regulated after mTOR inhibition (i.e., rapamycin, leucine, or glutamine deprivation) were enriched for MSI2 binding. These genes significantly overlap with three other distinct gene sets including energy metabolism, mRNA processing, and translation. In the "signaling and development module", we detected various pathways including gene sets related to "HSC versus CMP" and "Self-renewal", as well as other signaling pathways including TGFB1, RAS and MYC. Altogether, these results indicate sophisticated roles of MSI2 in regulating multiple critical cellular processes and pathways.

BIBLIOGRAPHY

1. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75,** 843–854 (1993).
2. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42,** D68–73 (2014).
3. Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19,** 92–105 (2009).
4. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173,** 20–51 (2018).
5. Yang, J. S. & Lai, E. C. Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. *Mol Cell* **43,** 892–903 (2011).
6. Song, J. J., Smith, S. K., Hannon, G. J. & Joshua-Tor, L. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305,** 1434–7 (2004).
7. Schirle, N. T. & MacRae, I. J. The crystal structure of human Argonaute2. *Science* **336,** 1037–40 (2012).
8. Schirle, N. T., Sheu-Gruttadauria, J. & MacRae, I. J. Structural basis for microRNA targeting. *Science* **346,** 608–13 (2014).
9. Jinek, M. & Doudna, J. A. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457,** 405–12 (2009).
10. Meister, G. Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet* **14,** 447–59 (2013).

11. Jee, D. *et al.* Dual Strategies for Argonaute2-Mediated Biogenesis of Erythroid miRNAs Underlie Conserved Requirements for Slicing in Mammals. *Mol Cell* **69,** 265–278 e6 (2018).

12. Hammell, C. M., Lubin, I., Boag, P. R., Blackwell, T. K. & Ambros, V. nhl-2 Modulates microRNA activity in Caenorhabditis elegans. *Cell* **136,** 926–38 (2009).

13. Schwamborn, J. C., Berezikov, E. & Knoblich, J. A. The TRIM-NHL protein TRIM32 activates microRNAs and prevents self-renewal in mouse neural progenitors. *Cell* **136,** 913–25 (2009).

14. Golden, R. J. *et al.* An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* **542,** 197–202 (2017).

15. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305,** 1437–41 (2004).

16. Meister, G. *et al.* Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15,** 185–97 (2004).

17. Jones-Rhoades, M. W., Bartel, D. P. & Bartel, B. MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol* **57,** 19–53 (2006).

18. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet* **16,** 421–33 (2015).

19. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455,** 64–71 (2008).

20. Hendrickson, D. G. *et al.* Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol* **7,** e1000238 (2009).

21. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466,** 835–40 (2010).

22. Eichhorn, S. W. *et al.* mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell* **56,** 104–15 (2014).

23. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508,** 66–71 (2014).

24. Schnall-Levin, M., Zhao, Y., Perrimon, N. & Berger, B. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3'UTRs. *Proc Natl Acad Sci U S A* **107,** 15751–6 (2010).

25. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).

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

27. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460,** 479–86 (2009).

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

29. Loeb, G. B. *et al.* Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol Cell* **48,** 760–70 (2012).

30. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153,** 654–65 (2013).

31. Grosswendt, S. *et al.* Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. *Mol Cell* **54,** 1042–1054 (2014).

32. Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., MacRae, I. J. & Joo, C. A Dynamic Search Process Underlies MicroRNA Targeting. *Cell* **162,** 96–107 (2015).

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

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

35. Urlaub, H., Hartmuth, K. & Lührmann, R. A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods* **26,** 170–81 (2002).

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

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

38. Meisenheimer, K. M. & Koch, T. H. Photocross-linking of nucleic acids to associated proteins. *Crit Rev Biochem Mol Biol* **32,** 101–40 (1997).

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

40. Konig, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* **13,** 77–83 (2012).

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

42. Bosson, A. D., Zamudio, J. R. & Sharp, P. A. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Mol Cell* **56,** 347–59 (2014).

43. Moore, M. J. *et al.* miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nat Commun* **6,** 8864 (2015).

44. Broughton, J. P., Lovci, M. T., Huang, J. L., Yeo, G. W. & Pasquinelli, A. E. Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Mol Cell* **64,** 320–333 (2016).

45. Wheeler, E. C., Van Nostrand, E. L. & Yeo, G. W. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *Wiley Interdiscip Rev RNA* **9** (2018).

46. Lu, Y. & Leslie, C. S. Learning to Predict miRNA-mRNA Interactions from AGO CLIP Sequencing and CLASH Data. *PLoS Comput Biol* **12,** e1005026 (2016).

47. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22,** 1173–83 (2012).

48. Nielsen, C. B. *et al.* Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13,** 1894–910 (2007).

49. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4** (2015).

50. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* **11,** R90 (2010).

51. Majoros, W. H. *et al.* MicroRNA target site identification by integrating sequence and binding information. *Nat Methods* **10,** 630–3 (2013).

52. Khorshid, M., Hausser, J., Zavolan, M. & van Nimwegen, E. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat Methods* **10,** 253–5 (2013).

53. Breda, J., Rzepiela, A. J., Gumienny, R., van Nimwegen, E. & Zavolan, M. Quantifying the strength of miRNA-target interactions. *Methods* **85,** 90–9 (2015).

54. Sonnenburg, S., Rätsch, G. & Rieck, K. in *Large Scale Kernel Machines* 73–103 (2007).

55. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6,** 26 (2011).

56. Evgeniou, T. & Pontil, M. Regularized multi-task learning. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004).

57. Tsochantaridis, I, Joachims, T, Hofmann, T & Altun, Y. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research* **6,** 1453–1484 (2005).

58. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27,** 2380–96 (2013).

59. Chi, S. W., Hannon, G. J. & Darnell, R. B. An alternative mode of microRNA target recognition. *Nat Struct Mol Biol* **19,** 321–7 (2012).

60. Wang, X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics* **32,** 1316–22 (2016).

61. Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I. & Hatzigeorgiou, A. G. Functional microRNA targets in protein coding sequences. *Bioinformatics* **28,** 771–6 (2012).

62. Sonnenburg, S., Zien, A., Philips, P. & Ratsch, G. POIMs: positional oligomer importance matrices–understanding support vector machine-based signal detectors. *Bioinformatics* **24,** i6–14 (2008).

63. Kedde, M. *et al.* RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell* **131,** 1273–86 (2007).

64. Kedde, M. *et al.* A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol* **12,** 1014–20 (2010).

65. Miles, W. O., Tschop, K., Herr, A., Ji, J. Y. & Dyson, N. J. Pumilio facilitates miRNA regulation of the E2F3 oncogene. *Genes Dev* **26,** 356–68 (2012).

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

67. Li, J. *et al.* Identifying mRNA sequence elements for target recognition by human Argonaute proteins. *Genome Res* **24,** 775–85 (2014).

68. Nolde, M. J., Saka, N., Reinert, K. L. & Slack, F. J. The Caenorhabditis elegans pumilio homolog, puf-9, is required for the 3'UTR-mediated

repression of the let-7 microRNA target gene, hbl-1. *Dev Biol* **305,** 551–63 (2007).

69. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* **79,** 351–79 (2010).

70. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455,** 58–63 (2008).

71. Nam, J. W. *et al.* Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol Cell* **53,** 1031–1043 (2014).

72. Erhard, F. *et al.* Widespread context dependency of microRNA-mediated regulation. *Genome Res* **24,** 906–19 (2014).

73. Lu, L. F. *et al.* A Single miRNA-mRNA Interaction Affects the Immune Response in a Context- and Cell-Type-Specific Manner. *Immunity* **43,** 52–64 (2015).

74. Vigorito, E., Kohlhaas, S., Lu, D. & Leyland, R. miR-155: an ancient regulator of the immune system. *Immunol Rev* **253,** 146–57 (2013).

75. Mashima, R. Physiological roles of miR-155. *Immunology* **145,** 323–33 (2015).

76. Eis, P. S. *et al.* Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc Natl Acad Sci U S A* **102,** 3627–32 (2005).

77. Seddiki, N., Brezar, V., Ruffin, N., Levy, Y. & Swaminathan, S. Role of miR-155 in the regulation of lymphocyte immune function and disease. *Immunology* **142,** 32–8 (2014).

78. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

79. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

80. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40,** 4288–97 (2012).

81. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11,** R25 (2010).

82. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

83. Kramer, N. J. *et al.* Altered lymphopoiesis and immunodeficiency in miR-142 null mice. *Blood* **125,** 3720–30 (2015).

84. Mildner, A. *et al.* Mononuclear phagocyte miRNome analysis identifies miR-142 as critical regulator of murine dendritic cell homeostasis. *Blood* **121,** 1016–27 (2013).

85. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43,** e47 (2015).

86. Cho, S. *et al.* miR-23 approximately 27 approximately 24 clusters control effector T cell differentiation and function. *J Exp Med* **213,** 235–49 (2016).

87. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22,** 2008–17 (2012).

88. Haasch, D. *et al.* T cell activation induces a noncoding RNA transcript sensitive to inhibition by immunosuppressant drugs and encoded by the proto-oncogene, BIC. *Cell Immunol* **217,** 78–86 (2002).

89. Van den Berg, A. *et al.* High expression of B-cell receptor inducible gene BIC in all subtypes of Hodgkin lymphoma. *Genes Chromosomes Cancer* **37,** 20–8 (2003).

90. O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. & Baltimore, D. MicroRNA-155 is induced during the macrophage inflammatory response. *Proc Natl Acad Sci U S A* **104,** 1604–9 (2007).

91. Ceppi, M. *et al.* MicroRNA-155 modulates the interleukin-1 signaling pathway in activated human monocyte-derived dendritic cells. *Proc Natl Acad Sci U S A* **106,** 2735–40 (2009).

92. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–8 (2011).

93. Wang, Y. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell* **25,** 69–80 (2013).

94. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495,** 333–8 (2013).

95. Piwecka, M. *et al.* Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science* **357** (2017).

96. Barrett, S. P. & Salzman, J. Circular RNAs: analysis, expression and potential functions. *Development* **143,** 1838–47 (2016).

97. Hausser, J., Syed, A. P., Bilen, B. & Zavolan, M. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res* **23,** 604–15 (2013).

98. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320,** 1643–7 (2008).

99. Gruber, A. R. *et al.* Global 3′ UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat Commun* **5,** 5465 (2014).

100. Pinzon, N. *et al.* microRNA target prediction programs predict many false positives. *Genome Res* **27,** 234–245 (2017).

101. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33,** 831–8 (2015).

102. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26,** 990–9 (2016).

103. Planas, A. P., Zhong, X. & Rayner, S. miRAW: A deep learning approach to predict miRNA targets by analyzing whole miRNA transcripts. *bioRxiv* (2017).

104. Kim, D. *et al.* General rules for functional microRNA targeting. *Nat Genet* **48,** 1517–1526 (2016).

105. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15,** 829–45 (2014).

106. Mukherjee, N. *et al.* Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol Cell* **43,** 327–39 (2011).

107. Jiang, P., Singh, M. & Coller, H. A. Computational assessment of the cooperativity between RNA binding proteins and MicroRNAs in Transcript Decay. *PLoS Comput Biol* **9,** e1003075 (2013).

108. Van Nostrand, E. L. *et al.* A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* (2017).

109. Ke, S. *et al.* A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev* **29,** 2037–53 (2015).

110. Mauer, J. *et al.* Reversible methylation of m(6)Am in the 5′ cap controls mRNA stability. *Nature* **541,** 371–375 (2017).

111. Benhalevy, D., McFarland, H. L., Sarshad, A. A. & Hafner, M. PAR-CLIP and streamlined small RNA cDNA library preparation protocol for the identification of RNA binding protein target sites. *Methods* **118-119,** 41–49 (2017).

112. Zarnegar, B. J. *et al.* irCLIP platform for efficient characterization of protein-RNA interactions. *Nat Methods* **13,** 489–92 (2016).

113. Daniels, D. L. *et al.* Examining the complexity of human RNA polymerase complexes using HaloTag technology coupled to label free quantitative proteomics. *J Proteome Res* **11,** 564–75 (2012).

114. McMahon, A. C. *et al.* TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* **165,** 742–53 (2016).

115. Xu, W., Rahman, R. & Rosbash, M. Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein. *RNA* **24,** 173–182 (2018).

116. Schmiedel, J. M. *et al.* MicroRNA control of protein expression noise. *Science* **348,** 128–32 (2015).

117. Denzler, R. *et al.* Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Mol Cell* **64,** 565–579 (2016).

118. Thomson, D. W. & Dinger, M. E. Endogenous microRNA sponges: evidence and controversy. *Nat Rev Genet* **17,** 272–83 (2016).

119. Park, S. M. *et al.* Musashi-2 controls cell fate, lineage bias, and TGF-beta signaling in HSCs. *J Exp Med* **211,** 71–87 (2014).

120. Rentas, S. *et al.* Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells. *Nature* **532,** 508–511 (2016).

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

122. Zhang, L. *et al.* Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing. *Elife* **4** (2015).

123. Okano, H. *et al.* Function of RNA-binding protein Musashi-1 in stem cells. *Exp Cell Res* **306,** 349–56 (2005).

124. Kawahara, H. *et al.* Neural RNA-binding protein Musashi1 inhibits translation initiation by competing with eIF4G for PABP. *J Cell Biol* **181,** 639–53 (2008).

125. Kharas, M. G. *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med* **16,** 903–8 (2010).

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