Understanding MicroRNA Targeting with High-Throughput Biochemistry

by

Sean E. McGeary

Sc.B., Biophysics (2009)
Brown University

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2021

Signature of author .....
Sean E. McGeary
Department of Biology
December 23rd, 2020

Certified by .....
David P. Bartel
Professor of Biology
Thesis Supervisor

Accepted by .....
Amy E. Keating
Professor of Biology and Biological Engineering
Co-Director, Biology Graduate Committee

Understanding MicroRNA Targeting with High-Throughput Biochemistry

by

Sean E. McGeary

Submitted to the Department of Biology on December 23, 2020
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Abstract

MicroRNAs (miRNAs) are short RNAs that, in complex with Argonaute (AGO) proteins, guide repression of mRNA targets. miRNAs negatively regulate most mammalian mRNAs, and disruption of this regulation often results in severe defects at the cellular and organismal level. miRNA repression occurs primarily through base-pairing between the miRNA seed region (nucleotides 2–8) and mRNA 3′-UTR sites, leading to transient recruitment of mRNA-destabilizing factors. However, only a small fraction of the gene-expression changes caused by a miRNA can currently be predicted, which precludes a deeper understanding of how miRNA regulation impacts the animal transcriptome.

miRNA targeting efficacy should in principle be a function of the affinity between AGO–miRNA complexes and their targets. However, only a few such measurements had been reported, with measured values differing from those predicted for RNA–RNA pairing in solution. We therefore adapted a high-throughput biochemical platform utilizing random-sequence RNA libraries to obtain the vast quantity of affinity values required to predict miRNA targeting efficacy. Through a novel analytical approach, we assigned relative dissociation (K_D) constants to all binding sites ≤12 nt in length, for six miRNAs. These analyses revealed unanticipated miRNA-specific differences in the affinity of similar sites, unique sites for different miRNAs, and a 100-fold influence of flanking dinucleotide context surrounding a site. These measurements informed a biochemical model of miRNA targeting that outperformed all existing models of miRNA targeting, which was extended to all miRNAs using a convolutional neural network (CNN) trained on both affinity and repression data.

We also applied this high-throughput biochemical approach to understand the role of the miRNA 3′ region using partially random RNA libraries. We found unique 3′-pairing preferences for each miRNA, and evidence for two distinct binding modes. The miRNA-specific differences and two binding modes depended on G nucleotides in the miRNA 3′ region, thus providing a heuristic by which to extend these findings to target prediction in vivo.

This work establishes high-throughput biochemistry combined with mathematical modeling and deep learning as a powerful paradigm for building quantitative models of gene regulation, which might aid in eventually building a complete model of the cell.

Thesis Supervisor: David P. Bartel
Title: Professor

Acknowledgements

Writing this dissertation nine month since the onset of the COVID-19 pandemic, and its accompanying societal changes, inspires heightened reflection regarding the community of people to whom I am indebted, for their care and generosity toward me as it pertains both to my professional career and to my continued development as a person. These acknowledgments represent my meager attempt to distill the meaning and inspiration that so many individuals have brought to my life.

It is truly challenging to put into words the gratitude I have for the role that my advisor, David Bartel, has played in my development as a thinker and scientist. Perhaps the greatest lesson I have learned under his supervision is that true understanding of anything is tantamount to the process of inspecting each individual detail. Perhaps the second greatest lesson is that scientific growth comes through continuous scrutiny of one's own ideas and assumptions. What I also should have learned, but didn't, is that writing shorter sentences makes them easier to understand. I am also indebted to Phillip Sharp and Christopher Burge for their support and critical assessment of my work over the years. I am thankful in particular to Christopher Burge for the opportunity to rotate with his lab, where I first contended with random-sequence binding experiments from a theoretical perspective, well before performing any such experiments myself. Thanks also to Phillip Zamore, for agreeing to serve as the outside member of my Thesis Committee. The rigorous experimental standard set by your research served as an example for my own biochemistry experiments. I also would like to thank Frank Solomon, for his willingness to provide honest perspective to trainees when they need it most.

The Bartel Lab has been a wonderful home to me over the years, as it has always been an environment that fosters critical scientific thinking and encourages engagement. Indeed, while the makeup of the lab as I depart is almost entirely different than when I began, there is a quality that has persisted, evident in the continued wide participation in Group Meeting, Idea Club, and Journal Club, that I hope remains a fixture of the lab into the future. I am grateful to Katrin Heindl both for instilling in me logistical rigor and for the candid presence she brought to every situation. I also feel indebted to Stephen Eichhorn, Alex Subtelny, Vikram Agarwal, David Weinberg, Olivia Rissland, and David Garcia, for the enumerable discussions that are part and parcel of my initial experimental and conceptual growth. I am indebted to Jeff Morgan, with whom I was bay-mates for the longest period of my own PhD, which unsurprisingly coincided with the entirety of Jeff's PhD. I am specifically thankful to Jeff for enabling our bay to be one in which the mutual respect was clear, wherein we never had to talk when we didn't want to, which made it all the richer when we did. I also have to thank Matt Getz and Ben Kleaveland, to whom I feel bonded over the many years we shared in the lab, having enumerable discussions of science and nonscience alike.

I have been deeply inspired by those individuals whom I have had the opportunity to see join the lab and grow into the great scientists they have become. I am thankful in particular for my relationship with Tim Eisen. His enthusiasm for science was evident in the quickness with which he engaged with each member of the lab on their research, and by how quickly he developed a singular command of the literature as it pertained to his project and those of others. I am also immensely thankful for my relationship with Charlie Shi. When considering how in the early days of his project we talked extensively about what he might do should no hit arise in his screen, I am so happy to see the success that his project has brought him, and am only excited to see where it leads.

I feel deep gratitude toward Kathy Lin; our collaboration has been the defining experience of my PhD. I am thankful for the innumerable ways our discussions have challenged me to think more deeply, and to specifically contend with how the realm of machine learning enables further discovery as it pertains to nucleic acid biology. I am also deeply thankful to Thy Pham, and am thrilled that the future of the research questions I pursued are in the hands of someone with such thoughtfulness and determination.

While in graduate school, I have made friends with truly incredible people, and had experiences with them that will remain with me for the rest of my life. The Family Dinners that dominated the social experience of the early portion of my graduate career were always something I looked forward to. Thank you to Jonathan Coravos, Doug Cattie, Cory Pender, and Mike Erickson, for making these nights so special. I also am deeply thankful that I had the opportunity to experience the force of nature that was Kotaro Kelley, whose enthusiasm and generosity were without limit.

I would be remiss to not acknowledge those friends of mine that predate the beginning of my graduate training, as their presence in my life has girded me and kept me accountable with respect to the outside world. Thank you, Josh Morrison, for the time while you were also at MIT, when you immediately included me in the world you were building in real time. Thank you also to Kevin Neal, whom I can go exceedingly long amounts of time without speaking to, and always find, as a testament to our relationship, that when we do finally talk that nothing has changed. I am also indebted to John Goodfellow, for our numerous discussions of science and music over the years, which always make me see things differently than I had previously. Thanks as well to Robert Smith III, for the warm, exciting times whenever you visited Boston. Lastly, I am thankful to Ben Mandel, Daniel Lennard, Gerry Bell, and Andrew Underberg, for our COVID19-era weekly film club. In these strange times, the routine of watching a movie and then discussing it over Zoom has been a bedrock of normalcy that I have so deeply appreciated. I hope it continues as we move toward a brighter future.

I am fortunate in that I have been able to spend the latter half of my graduate career living with two incredible roommates, Alex Godfrey and Kunle Demuren. I'm furthermore thankful to Bridget Begg, Aaron Hosios, and Steve Sando, for making the days of lockdown less hard than they otherwise would have been. In particular, I thank Bridget for always believing in me. And I absolutely have to thank Alex, for a friendship that has filled my PhD experience with so much joy and meaning. You, and the bond we share, inspire me to try to be my best self, and for that I can't thank you enough.

I am incredibly fortunate to have grown up in a family in which love and support were in limitless supply. Danielle, I am so proud of the strong woman that you have become. Also, thanks for taking care of your little brother, whenever he needs it. Anthony, I am so glad you are a part of the family, and I look forward to every time we see each other. Ansley becomes more wonderful with each passing day, and I can't wait until she's old enough to talk about microRNAs. Finally, I would like to thank my parents. Mom and Dad, I will never be able to express how grateful I am to you, for teaching me the importance of education, for deeply instilling in me right versus wrong, and enabling me to have the confidence to always believe in myself. Without you, this thesis, and so much else that has happened in my life, would never have been possible.

Table of contents

Chapter 1. Introduction

The regulation of gene expression

The essential organizing principle of biological science is the pursuit of increased understanding of the organisms that populate the planet. This field of study is currently dominated by research at the molecular scale, owing in large part to foundational experiments performed during the middle of the 20th century elucidating the material basis of heredity (Hershey and Chase, 1952; Watson and Crick, 1974; Zamenhof et al., 1952). Indeed, the contemporary model for how cells participate in the diverse processes collectively referred to as "life" is nearly identical to that which emerged more than 50 years ago (Crick, 1970), whereby 1) genes correspond to contiguous informational segments within chromosomal DNA, 2) these segments can be used to generate RNA (and protein) molecules with defined biochemical functions, and 3) the co-occurring functions of each of the expressed gene products necessarily mediates cellular physiology. The unifying aim of molecular biology is therefore to understand the nature of each gene, which includes both identifying the role of each gene in the context of cellular-to-organismal physiology and developing a descriptive understanding of the molecular mechanisms by which each gene operates.

The importance of which genes are expressed in determining the functional state of a cell or tissue is borne out by countless studies reporting variation in the abundance of individual messenger RNAs (mRNAs) and/or proteins between different biological systems, including but not limited to reports comparing different mouse neuronal cell types (Lein et al., 2007), different mouse and human tissues (Huttlin et al., 2010; Kim et al., 2014), and different stages within the yeast cell cycle (Spellman et al., 1998). Furthermore, large-scale analysis of genetic variation among individuals has implicated >3000 genes as being haploinsufficient, due to the near-

complete depletion of loss-of-function alleles for these genes among the >60,000 individuals comprising the study (Lek et al., 2016). The prevalence of haploinsufficiency, and as well the pleiotropic, deleterious fitness consequences of whole-chromosome aneuploidy for most organisms (Siegel and Amon, 2012), together underscore the importance of maintaining the cellular abundance of many expressed genes within a particular range, rather than merely turning some genes "on," and others "off."

Consistent with this picture, upwards of 1/3rd of the human genome exhibits functional potential, as determined by variation in accessibility across 125 distinct cell types measured through DNase I hypersensitivity assays (Thurman et al., 2012). By comparison, no more than 3% of the human genome likely encodes a functional polypeptide, suggesting that the vast majority of the genomic sequence contained within these variably accessible regions functions to establish and maintain gene-regulatory mechanisms (Thurman et al., 2012). Indeed, the regulation of RNA transcription, necessarily the first step in functional gene expression, is well-established: hundreds of distinct transcription factors bind throughout the genome to noncoding "promoter" elements directly 5′ of the transcriptional start sites of individual genes, and as well as to "enhancer" elements which can lead to larger regulatory changes for genes positioned arbitrarily large distances away within the genome (Gasperini et al., 2020). Regulation of animal transcription rates by these and other mechanisms including chemical modification of histone proteins, chemical modification of the DNA itself, and formation of sub-nuclear environments called topologically associated domains (Pombo and Dillon, 2015) has been reported as contributing as much as 73% of the variation in overall protein levels, as calculated from the re-analysis of studies performing paired proteomics and time-resolved RNA-seq measurements in mouse NIH3T3 cells (Li et al., 2014; Schwanhäusser et al., 2011). These studies both

corroborate the importance of transcriptional control and also indicate that at least one quarter of gene expression control occurs post-transcriptionally[1], through the regulation of mRNA degradation, mRNA translation, and protein degradation.

Evidence of regulated mRNA stability can be found in studies of the so-called "immediate early" genes[2] such as c-*fos* and c-*jun* (Sheng and Greenberg, 1990), whose mRNAs exhibit pulsatile induction kinetics contributed to by both rapid transcription (Bartel et al., 1989), and rapid turnover characterized by a half-life of 10–15 minutes (min) (Sheng and Greenberg, 1990), in comparison to reported median half-lives of 2–9 hours (h) drawn from global metabolic labeling studies in NIH3T3 cells (Eisen et al., 2020a; Schwanhäusser et al., 2011). The increased instability of the mRNAs of these genes in comparison to that of others was linked to A/U-rich elements within the translated open reading frame (ORF) and 3′ untranslated region (UTR) within the RNA molecule (Hentze, 1991). Around this time, a sequence was also identified within the 5′ UTR of the *ferretin* mRNA that formed a stem loop and conferred iron-dependent translational regulation of the mRNA (Hentze et al., 1987). By analogy to the noncoding-but-functional DNA sequence elements within the genome, a picture was beginning to emerge that the noncoding sequences up- and down-stream of the genic coding sequence of an mRNA served to modulate both the rate of its translation and the time until its eventual degradation, through specific sequences that associate with known and yet-unknown trans-acting protein factors.

[1]Here "post-transcriptionally" is defined to mean occurring after the completion of transcription *and* any processing of the RNA into its final form. This is because the time-resolved RNA sequencing measurements from which the reported percentages were derived came from reads mapped to the spliced, fully processed sequence. Indeed, the complex mechanisms by which the splicing of Pol II transcription products is regulated are not addressed in this work, nor are the biogenesis and function of rRNA, snRNA, and tRNA molecules.

[2]The naming of these genes comes from the observation that those genes that responded to trans-synaptic stimulation and/or membrane electrical activity in neurons fell into two broad categories. Those whose transcription began rapidly and transiently upon stimulation were called immediate early genes, and those whose response was slower and more persistent were named late response genes.

Discovery of small RNAs

The very first evidence of what would eventually be called RNA interference (RNAi) came from experiments initially intending to increase the purple coloration of petunias (Napoli et al., 1990): transformation of these flowers with a transgene encoding the pigment-producing enzyme chalcone synthase caused the unanticipated whitening of the petals, rather than darkening their hue. The molecular nature underpinning this phenomenon[3], termed "co-suppression," was mysterious, as there was no precedent for increased genomic copy number of a gene leading to loss of its expression. Around the same time, efforts to silence expression of *unc-22* and *unc-45* by injection of antisense RNA into the roundworm *Caenorhabditis elegans* (*C. elegans*) were proving successful (Fire et al., 1991). This was thought to be caused by the antisense RNA hybridizing with *unc* mRNA, thereby preventing its translation. In a later study, germline-injection of antisense RNA derived from cDNA was used to confirm that the identity of the cDNA was, in fact, the embryonic polarity-promoting gene *par-1* (Guo and Kemphues, 1995). However, the authors found that injection of either the sense or the antisense RNA caused a similar percentage of developmental arrest upon germ-line injection, complicating the interpretation that the antisense RNA was suppressing gene function by hybridizing directly with the mRNA.

These mysteries were eventually clarified, with the first advance coming from further experiments in *C. elegans*, whereby it was determined that long (i.e., several-hundred-nt) double-stranded RNA (dsRNA) was the relevant trigger for RNAi, and further that the low amount of dsRNA required for RNAi rendered a direct-hybridization model yet more implausible (Fire et al., 1998). The requirement for dsRNA over sense or antisense RNA for RNAi was also observed

[3]The authors posit in the discussion that "the erratic and reversible nature of the CHS transgene effect suggests the involvement of methylation."

using cell-free systems developed from syncytial blastoderm *Drosophila* embryos, which both confirmed the requirements for activation of this regulatory mechanism, and as well provided a powerful system with which to further its study (Tuschl et al., 1999). Indeed, this in vitro system enabled the discovery that RNAi occurred through the dsRNA being processed to staggered 21–23 nt fragments that in turn served to guide ATP-dependent endonucleolytic cleavage, or "slicing," of the targeted mRNA (Zamore et al., 2000). Shortly thereafter, efficient RNAi was demonstrated directly using 21- or 22-nt RNA duplexes with 5′ hydroxyls[4], 3′ hydroxyls, and 2-nt 3′ overhangs (Elbashir et al., 2001a), which occurred without the production of the staggered 21–23 nt RNA fragments observed when introducing long dsRNA, and enabled productive RNAi in mammalian cell culture (Elbashir et al., 2001b). A further advantage of these duplex was that the slicing was positionally defined, occurring at the phosphodiester bond linking the target nucleotides pairing to nucleotides 10 and 11 of the complementary RNA within the duplex. Since these ~21-nt RNAs were almost certainly the direct effector molecules of RNAi, they were named short interfering RNAs, or siRNAs (Elbashir et al., 2001a).

Contemporaneous with the studies determining the molecular nature of RNAi, unbiased screens conducted in *Caenorhabditis elegans* for heterochronic (i.e., important at distinct stages in development) genes identified two loci bearing unprecedented molecular characteristics: *lin-4* (Ambros, 1989; Lee et al., 1993; Wightman et al., 1993) and *let-7* (Reinhart et al., 2000). These two negative-regulatory genes were surprising in that each of their ultimate functional products was not a protein, but rather a 21- or 22-nt RNA. Additionally, these two RNAs exhibited imperfect complementarity to sites within the 3′ UTRs of their downstream regulatory target

[4]A guide RNA requires a 5′ phosphate in order to be loaded into an Argonaute protein. However, the synthetic duplexes did not require this modification because of the presence of an endogenous 5′-kinase activity in the lysates (Elbashir et al., 2001a) and cell culture models (Elbashir et al., 2001b) used in both studies.

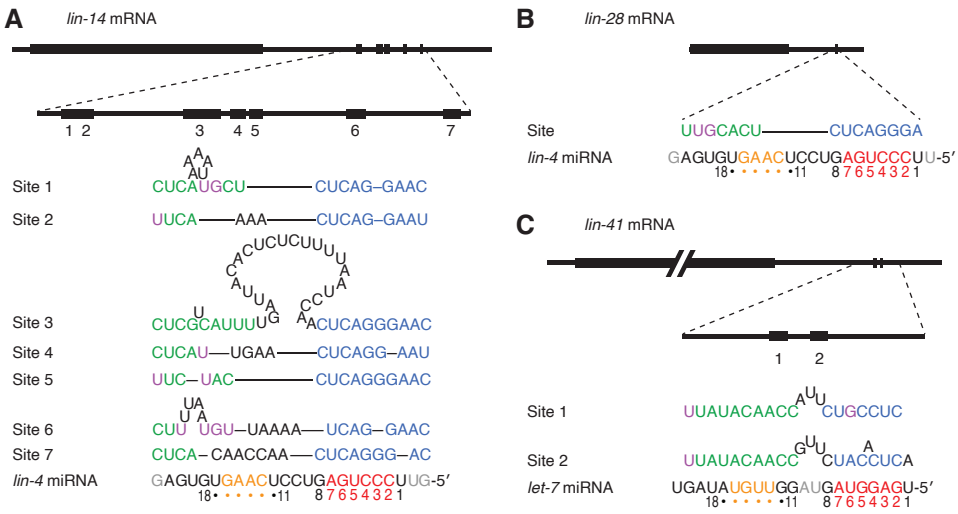


Figure 1. First identified miRNAs and their targets.
(**A**–**C**) Schematics depicting the mRNA, the 3′-UTR site sequences, and miRNA sequence for *lin-14* repression by *lin-4* (A), *lin-28* repression by *lin-4* (C), and *lin-41* repression by *let-7*, as reported in Lee et al. (1993) and Wightman et al. (1993) (A), Moss et al. (1997) (B), and Reinhart et al. (2000) (C), respectively. For each target site, the proposed site architecture is displayed showing both seed pairing (blue) and 3′ pairing (green), with wobble pairs indicated (purple). The gray nucleotides on either side of the *lin-4* sequence in (A) and (B) reflect the uncertainty of the end definition of the mature miRNA sequence at the time of publication.

genes, *lin-14* and *lin-28*, in the case of *lin-4*, and *lin-41*, in the case of *let-7* (Lee et al., 1993; Moss et al., 1997; Reinhart et al., 2000; Wightman et al., 1993). RNase-protection assays performed with *lin-14* indicated that the protein, and not the mRNA levels, were downregulated by the *lin-4* gene, suggesting that these heterochronic small RNAs functioned by binding to the 3′-UTR sites and promoting translational repression of their target genes.

let-7 differed from *lin-4* in that its full (i.e., 21-nt) sequence was found to be conserved across a diversity of metazoan species inclusive of flies, molluscs, and vertebrates (Pasquinelli et al., 2000). The appreciation that these short RNA species were not merely an idiosyncrasy of early *C. elegans* development, as well as the finding that synthetic siRNAs could function in human cells (Elbashir et al., 2001b), motivated the design of small RNA cloning and sequencing approaches, in order to profile the diversity of sRNAs in animal cells (Elbashir et al., 2001a;

Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). The studies in total reported 16 *Drosophila*, 55 *C. elegans*, and 21 human examples of these 21–23-nt sRNAs, naming them microRNAs (or miRNAs) due to both their short size, and their as-yet unclear role in animal cells.

Molecular modes of miRNA-mediated repression

The contemporary understanding of miRNA-mediated repression, and indeed its distinction from RNAi, is extensive (Bartel, 2018). miRNAs constitute a class of 21–23-nt small RNAs that are processed from hairpin precursors and loaded into Argonaute (Ago) proteins (Liu et al., 2004). miRNAs are a feature of both plant and animal genomes, although the two pathways have completely different miRNA sequences, and distinct mechanisms of biogenesis and repression, suggesting the pathway either evolved separately, or that the plant, animal, or both pathways diverged considerably from that which was present in last common ancestor of plants and animals (Moran et al., 2017). The widespread biological importance of animal miRNAs is evident from mouse knock-out studies—removal of some or all members of at least 20 miRNA families (i.e., all miRNAs with an identical sequence at positions 2–8) conserved throughout bilateria results in significantly deleterious phenotypes[5] (Bartel, 2018), with nine causing lethality (Dooley et al., 2015; Han et al., 2015; Heidersbach et al., 2013; Liu et al., 2008; Penzkofer et al., 2014; Sanuki et al., 2011; Shibata et al., 2011; Song et al., 2014; Wang et al.,

[5]In contrast to results from mice, knockout or disruption of individual miRNA genes in *C. elegans* predominantly resulted in no phenotype (Miska et al., 2007). This in principle could be due to functional redundancy, owing to the >60% of *C. elegans* miRNAs sharing a seed sequence with at least one other miRNA. A later study generating worms knocked out for all paralogs of individual seed families found strong defects for only three of the 15 families tested, ruling out seed redundancy as the primary explanation for the lack of phenotypes (Alvarez-Saavedra and Horvitz, 2010). It has since been argued that the difference in phenotypic consequence upon miRNA loss between worms and mice is due to the increased tolerance of worms to abnormalities in their differentiated cells in comparison to mammals, since many of the mouse lethality phenotypes in mice occur very late in or after development of the body plan (Dexheimer et al., 2020).

2017; Wei et al., 2014) and the remainder exhibiting effects as diverse as infertility (Ahmed et al., 2017; Hasuwa et al., 2013; Shibata et al., 2011), intestinal hypertrophy (Madison et al., 2013), altered liver regeneration (Wu et al., 2015), reduced lifespan (Smith et al., 2012), resilience to stress (Andolina et al., 2016), sensory hair-cell degeneration (Kuhn et al., 2011; Lewis et al., 2009; Mencía et al., 2009), and myelination defects (Wang et al., 2017). There are ~500 stringently annotated human miRNA genes, which amounts to 1–3% of human genes[6] (Bartel, 2018; Kozomara et al., 2019; Pertea et al., 2018), and there is evidence that >60% of human mRNAs harbor a miRNA site with signal for conservation greater than that of its surrounding sequence context (Friedman et al., 2009). These findings together underscore the centrality of the role miRNAs play in cellular gene-regulatory control.

Because the association between Ago and a miRNA is a stable interaction that typically persists for many hours to days (Guo et al., 2015; Kingston and Bartel, 2019; Rooij et al., 2007), these complexes are, in effect, modular RNA binding proteins (RBPs), with binding specificity conferred by the loaded miRNA, rather than by a constitutive domain of the protein itself. Ago-miRNA complexes elicit repression by associating with binding sites located primarily in mRNA 3′ UTRs that minimally contain perfect complementarity to miRNA nucleotides 2–7, known as the miRNA seed (Bartel, 2009; Lai, 2002; Lewis et al., 2003, 2005). Many sites additionally have pairing to miRNA nucleotide 8, a target A nucleotide across from miRNA nucleotide 1, or both, with either feature further increasing the efficacy of repression (Bartel, 2009; Grimson et al., 2007; Lewis et al., 2005). The consistency of these four possibilities has led to the classification of "canonical" miRNA sites, with the 8-nt site known as the 8mer, the two 7-nt

[6]This number is 2.5% if using the rule-of-thumb of 20,000 human genes, and 1.25% if including newer estimates of ~20,000 coding transcripts and ~22,000 noncoding transcripts assembled from RNA-seq collected in the Genotype-Tissue Expression (GTEx) project (Pertea et al., 2018).

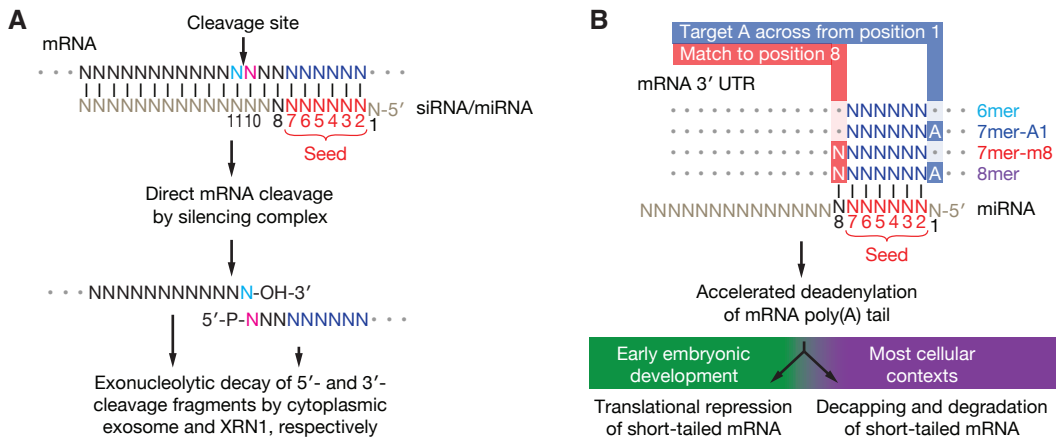


Figure 2. Molecular mechanisms of small RNA–based repression.
(**A**) Endonucleolytic cleavage of target mRNAs directed by siRNAs and miRNAs. The target mRNA is cleaved at sites of perfect or near-perfect complementarity to the small RNA guide sequence, across from positions 10 (pink) and 11 (cyan) of the small RNA. This allows both fragments to be rapidly degraded by cytoplasmic endonucleases. While all targeting by siRNAs definitionally occurs through direct cleavage, only a handful of endogenous miRNA targets are repressed by this mechanism. (**B**) Repression by most miRNAs, which occurs with shorter regions of complementarity to the miRNA seed (red). The four founding canonical sites are represented, which all contain full complementarity to the seed and possibly a match to miRNA position 8, an A across from miRNA position 1, or both. Pairing at such a site causes more rapid deadenylation of the mRNA, which eventually leads to translational repression or mRNA destabilization, depending on the cellular context.

sites known as the 7mer-m8 and 7mer-A1, and the site pairing only to nucleotides 2–7 known as the 6mer (Bartel, 2009; Lewis et al., 2005).

miRNAs decrease the amount of protein produced by their mRNA targets through a combination of mRNA destabilization and translational repression (Bazzini et al., 2012; Eichhorn et al., 2014; Eisen et al., 2020b; Guo et al., 2010; Jonas and Izaurralde, 2015; Subtelny et al., 2014). The extent to which each of these modes occurs is a function of the cellular context: in early development prior to zygotic genome activation in zebrafish (and presumably most miRNA-containing animals), translational repression is the predominant mode of repression (Bazzini et al., 2012; Subtelny et al., 2014), whereas in adult tissues and most cell culture models observed at steady state, mRNA destabilization predominates (Baek et al., 2008; Eichhorn et al.,

2014; Guo et al., 2010). Indeed, determining the amount of each of these two modes of repression has been a major subject of inquiry over the past decade, and has been greatly aided by the ability to perform global measurements of mRNA expression levels, through RNA-seq, and global measurements of ribosome engagement, through ribosome footprint profiling (Bazzini et al., 2012; Guo et al., 2010; Ingolia et al., 2009).

A description of the molecular nature of miRNA-mediated mRNA destabilization (and of translational repression within early embryonic contexts) is found in the next section, followed by a shorter section describing RNAi, and the rare circumstances in which Ago–miRNA complexes participate in RNAi-like silencing rather than the more common form of miRNA-mediated repression[7].

The role of the mRNA poly(A) tail in miRNA-mediated repression

miRNAs predominantly exert their destabilizing effect on mRNAs by accelerating the rate at which mRNAs proceed through their normal life cycle (Eisen et al., 2020b). Eukaryotic mRNAs harbor a 7-methylguanosine cap connected by a 5′–5′ phosphate linkage at their 5′ ends (Sonenberg et al., 1978) and an untemplated poly(A) tail at their 3′ ends (Rosenthal et al., 1983), which both can serve to promote mRNA stability and translation (Goldstrohm and Wickens, 2008; Weill et al., 2012). The poly(A) tail is added during the process of transcriptional termination; almost every animal mRNA contains a cleavage-and-polyadenylation signal sequence within its 3′ UTR that, upon nascent transcription of this sequence element by still-processing Pol II, signals for endonucleolytic cleavage at that site, followed by enzymatic

[7]No attempt is made to suggest target slicing by a miRNA does not constitute miRNA-mediated repression, since the miRNA is still repressing protein output. However, the near-ubiquity with which non–RNA biologists conflate RNAi and miRNA-mediated repression has motivated some attempt within this thesis to emphasize that slicing-based targeting is relevant to only an extreme minority of animal miRNA sites.

addition of ~200 nt of untemplated adenosines (Proudfoot et al., 2002). The cleavage-and-polyadenylation reaction also seems to be the mechanism by which the pre-mRNA is liberated from the still-transcribing locus, which serves as an incipient cue to terminate transcription that is eventually transduced to Pol II (Connelly and Manley, 1988; Logan et al., 1987; Proudfoot, 1989).

The poly(A) tail and 5′ cap together imbue the mRNA with the property of non-exponential decay. Exponential decay is characterized by all members of a population (in this case, of molecules) experiencing decay as a unitary, absolute process, with a probability of occurrence that is constant over time. Exponential decay is therefore "memoryless," in which neither the state of the molecule, nor its having persisted for more or less time, has any impact on the likelihood of the molecule's immediate, complete degradation. mRNAs are not well described by this regime because the poly(A) tail serves as a molecular timer, whereby cytoplasmic deadenylases PAN2–PAN3 and CCR4–NOT cause the gradual shortening of the poly(A) tail of individual molecules over time (Chen and Shyu, 2011; Meyer et al., 2010). Once the tail length has been shortened to ~20 nt on average, this molecular information is transduced to the 5′ end, leading to the decapping of the mRNA by the decapping complex DCP1–DCP2 (Chowdhury et al., 2007). Upon decapping, the mRNA is rapidly degraded, primarily by the cytoplasmic exonuclease XRN1 (Chen and Shyu, 2011).

miRNAs influence this degradation pathway by stimulating increased deadenylation of the poly(A) tail (Braun et al., 2012; Eisen et al., 2020b; Giraldez, 2006; Subtelny et al., 2014), by association of the AGO–miRNA complex with mostly unstructured proteins of the GW182 family (Braun et al., 2011, 2013; Eulalio et al., 2008). Because GW182 proteins interact with both the PAN2–PAN3 and CCR4–NOT complexes (Behm-Ansmant et al., 2006), miRNAs are

able to promote the more efficient shortening of poly(A) tails while bound to their targets. This causes, over the course of one, two, or perhaps many binding and dissociation events, an mRNA target to have a shorter poly(A) tail than it would in the absence of miRNA binding, such that it more quickly reaches the 20-nt tail-length threshold associated with rapid decapping and degradation (Cao and Parker, 2001; Eisen et al., 2020b, 2020a).

There is evidence that some amount of translational repression can occur due to the recruitment of the RNA helicase DDX6 by CCR4–NOT (Chen et al., 2014). DDX6 is known to promote decapping, which may lead to translational repression in the time between the initiation of decapping and the full degradation of the mRNA. However, in the only biological context with transcriptome-wide measurements demonstrating that miRNAs predominantly act through translational repression, this influence on translation comes from the same deadenylation-promoting activity of miRNA binding as described above (Subtelny et al., 2014). This difference in the ultimate effect of miRNA targeting is due to short-tailed mRNAs being translationally repressed, rather than degraded, in the early embryo. These insights also likely apply to the early frog and fly embryo, as both contexts establish a similar coupling between poly(A) tail–length and translational efficiency (Eichhorn et al., 2016; Subtelny et al., 2014); however, no direct measurements of mode of miRNA-mediated repression have been made in either system. Indeed, the early embryonic samples in which poly(A) tail–length and translation are coupled are developmental stages in which zygotic genome activation has either not yet or just occurred. Translational repression in this context is probably more desirable than mRNA destabilization, as it enables regulation of overall protein output without partial destruction of the transcriptome before it can be replaced by transcription of the zygotic genome (Eichhorn et al., 2014, 2016; Subtelny et al., 2014).

Rare RNAi-like repression of some animal miRNA targets

In some cases, miRNA complexes perform an alternative type of mRNA destabilization, whereby the Ago–miRNA complex catalyzes the endonucleolytic cleavage of its target RNA (Davis et al., 2005; Shin et al., 2010; Yekta et al., 2004), in a reaction that is chemically identical to that of RNAi. That a miRNA could perform RNAi was first shown with the demonstration that let-7 could direct efficient, multiple-turnover slicing of synthetic target RNAs in human cell extracts (Hutvágner and Zamore, 2002). This type repression requires both the extensive complementarity of the guide- and target-RNA (Becker et al., 2019; Elbashir et al., 2001a; Haley and Zamore, 2004; Wee et al., 2012), and an Ago protein capable of directing cleavage. Indeed, cleavage-competent Ago proteins are found in all domains of life, and maximum-likelihood phylogenetic trees constructed for both prokaryotic and eukaryotic Argonautes suggest nucleic acid–directed slicing was the ancestral role of this protein family (Swarts et al., 2014). However, only human Ago2 (AGO2) is strongly cleavage-competent among the four human paralogs (Liu et al., 2004; Meister et al., 2004; Rivas et al., 2005). This activity is presumably selectively maintained in part by the handful of highly-complementary miRNA targets, such as the highly complementary miR-196 site in the *Hoxb8* mRNA, which is active during limb development (Yekta et al., 2004). However, there is also evidence that the most consequential slicing activity of AGO2 occurs during the atypical biogenesis of two miRNAs important for normal erythroid development, miR-451 and miR-486 (Cheloufi et al., 2010; Chen et al., 2017; Cifuentes et al., 2010; Jee et al., 2018; Kretov et al., 2020). In any case, the remainder of this introduction will concern itself with those animal miRNA target sites whose repression proceeds through deadenylation, rather than direct slicing by the Ago–miRNA complex, as these are the sites

through which animal miRNAs predominantly exert their biological functions (Friedman et al., 2009).

Quantitative prediction of cellular miRNA targeting

The molecular mechanisms by which miRNAs repress their cellular targets appear to be well-established—that is, the list of proteins demonstrated to be important within the pathway[8], and the apparent modes by which they interact with the targeted mRNA and each other, has not undergone any substantive revision in recent years (Bartel, 2018; Jonas and Izaurralde, 2015). One notable exception to this is the discovery of a phosphorylation cycle acting directly on target-bound Ago–miRNA complex, mediated in humans by the kinase CSNK1A1 and the ANKRD52–PPP6C phosphatase complex (Golden et al., 2017), wherein disruption of the cycle impedes the efficacy of targeting (Golden et al., 2017; Huberdeau et al., 2017).

Even absent a complete understanding of their mechanism of action, the question remains as to how miRNAs exert their biological functions at the cell, tissue, and organismal scale. Since miRNAs function at the molecular level by directing mRNA repression throughout the transcriptome, this question is tantamount to understanding, upon expression of a particular miRNA, which mRNAs will be targeted by that miRNA, and the magnitude of the effect for each. Indeed, while the identification of the miRNA seed (Lewis et al., 2003), and the further establishment of canonical site types (Bartel, 2009; Brennecke et al., 2005; Lewis et al., 2005) constituted major advances in identifying which sites might be effective, they are not sufficient to quantitatively explain the effects of miRNAs: there many instances of seed site–harboring

[8]This statement refers specifically to the mRNA destabilization and translational repression mechanisms discussed in the prior sections. It does not refer to the miRNA biogenesis pathway, nor to any miRNA degradation pathways, including of target RNA–directed miRNA degradation (TDMD).

mRNAs that are not miRNA-responsive, and many instances of mRNAs without a site that are responsive (Grimson et al., 2007). This is consistent with results from analysis of 3′-UTR sequence evolution: highly-expressed, cell type–specific mRNAs have tended to avoid sites to co-expressed miRNAs, possessing on average 50% fewer 7mer sites to these miRNAs in comparison to control mRNAs (Farh et al., 2005; Stark et al., 2005). This is consistent with the notion that approximately 50% of such sites are mediating effective repression, as a function of variation in contextual features extrinsic to each site.

Indeed, a number of features have been identified that modulate the efficacy of a miRNA site. These include the total abundance of target sites to the given miRNA (where increased abundance leads to dilution of the miRNA among all of those sites, thereby weakening repression) (Garcia et al., 2011), the predicted stability with which the miRNA seed region will pair with its Watson–Crick complementary sequence (Garcia et al., 2011; Ui-Tei et al., 2008), the predicted stability with which the 3′-UTR sequence will form secondary structure occluding the linear target site (thereby decreasing the efficacy of the site) (Tafer et al., 2008; Wan et al., 2014), the local AU content near the site (Grimson et al., 2007; Nielsen et al., 2007), additional pairing to the miRNA 3′ end (Brennecke et al., 2005; Grimson et al., 2007), the preferential conservation of a site (Brennecke et al., 2005; Friedman et al., 2009; Krek et al., 2005), the distance of a site from either end of the 3′ UTR (Gaidatzis et al., 2007; Grimson et al., 2007; Majoros and Ohler, 2007), and the lengths of both the mRNA ORF and 3′ UTR (Agarwal et al., 2015; Hausser et al., 2009). These and a few other features have been integrated into a model predicting miRNA target gene expression, providing unambiguous improvement over when considering site type alone (Agarwal et al., 2015), and also outperforming myriad alternative computational approaches, with some informed by the predicted stability (i.e., the ΔG) of pairing

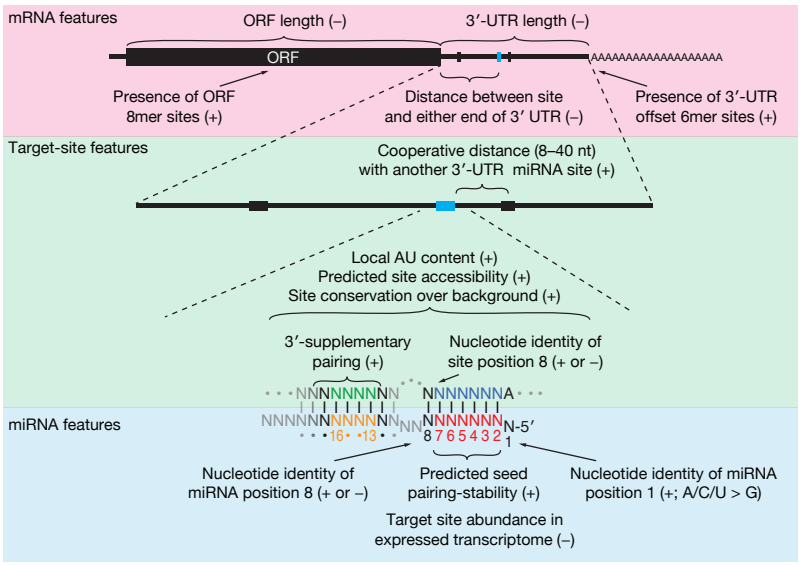


Figure 3. Features leading to quantitative differences in miRNA target site efficacy. Depicted are the 14 features utilized in Agarwal et al. (2015), as well as one more feature related to the cooperative spacing of miRNA sites (Grimson et al., 2007). "+" indicates that the feature leads to increased repression, "–" leads to decreased repression, and "+ or –" indicates that the effect of the feature depends on the site type.

between the full miRNA and target sequence (Anders et al., 2012; Gumienny and Zavolan, 2015; Krek et al., 2005) or from crosslinking and immunoprecipitation of Ago–miRNA complexes followed by high-throughput sequencing (CLIP-seq) (Khorshid et al., 2013). However, even in this context, only 16% of the transcriptome-wide effects of a miRNA could be explained (Agarwal et al., 2015), indicating either a significant gap in our understanding of the features relevant to miRNA targeting efficacy, or alternatively that the true signal from miRNA-mediated repression is small in comparison to both experimental noise and the secondary effects caused by repression of primary targets.

In addition to incomplete understanding of the effects of sequence context on site efficacy, another explanation for why quantitative modeling of miRNA-mediated repression has not achieved greater overall success is that target sequences other than the four canonical 6mer, 7mer-A1, 7mer-m8, and 8mer sites mediate functional repression (Hausser and Zavolan, 2014),

and that the omission of these noncanonical-yet-functional sites causes an under-estimation of predicted repression of some or many miRNAs, potentially missing some target mRNAs bearing only noncanonical sites. Indeed, the definition of a canonical site has itself expanded over time: the two 6-nt sites that are offset from the canonical 6mer by one nucleotide in either the 5′ (known as the 6mer-A1) (Jan et al., 2011; Kim et al., 2016) or 3′ direction (known as the offset 6mer or 6mer-m8) (Friedman et al., 2009) have more recently been considered canonical sites due to their frequent, if not ubiquitous, signal for repression in vertebrates. In addition, it has been appreciated since the identification of the *let*-7 sites in the 3′ UTR of the *C. elegans lin-41* mRNA that in some cases mismatched or bulged target nucleotides are tolerated within the seed, if sufficiently compensated for by extended pairing to the miRNA 3′ end, referred to as a 3′-compensatory site[9] (Brennecke et al., 2005; Reinhart et al., 2000). Also later identified were 11–12-nt sites with pairing beginning at miRNA position 4 or 5, referred to as centered sites (Shin et al., 2010). While both of these noncanonical site types have been detected with multiple miRNAs, thereby validating their function, they are rare within animal transcriptomes, comprising ~1% of preferentially conserved sites in human transcriptomes (Friedman et al., 2009), suggesting that the omission of these or of any yet-unknown, equally-rare noncanonical sites is not the predominant cause for the low performance of target prediction efforts.

The identification of both the canonical and noncanonical sites thus described comes from evidence of their function for multiple different miRNAs. If each miRNA did bind to a

[9]Of interest are the apparent evolutionary pressures acting on the two *let-7a* sites in the *lin-41* 3′ UTR that led to their 3′-compensatory site architecture—namely, the dual pressure to enable efficient targeting and repression by let-7a (Pasquinelli et al., 2000; Reinhart et al., 2000) while also not acquiring a seed-complementary site that would cause repression by other seed-family members earlier in larval development (Brancati and Großhans, 2018). Indeed, the much greater information content required to achieve repression using 3′-compensatory pairing compared to that of a canonical site indicates that 3′-compensatory sites, when found, might lead to especially strong organismal phenotypes when disrupted, even if the average change in target expression caused by the two site types were of similar magnitude.

distinct set of functional noncanonical sites, then elucidating these sets of sites on a per-miRNA basis would constitute an important advance in the understanding of miRNA targeting, as quantitative models of predicted targeting efficacy could be updated to include these distinct site profiles. While one could in principle look for miRNA-specific noncanonical sites directly within data generated from in vivo experiments such as miRNA transfection followed by RNA-seq, it would be challenging to disentangle which of the miRNA-specific *k*-mers that correlate with repression were due to direct association with the Ago–miRNA complex, and which were false-positives with the particular set of mRNAs for which repression was observed.

To this end, a variety of studies have generated compelling evidence of miRNA-specific noncanonical sites through the use of CLIP-seq (Chi et al., 2009, 2012; Grosswendt et al., 2014; Hafner et al., 2010; Hausser et al., 2009; Lipchina et al., 2011; Loeb et al., 2012). An extension of this protocol was later developed, enabling the crosslinking, ligation, and sequencing of hybrids (CLASH) (Helwak et al., 2013; Kudla et al., 2011), which generates chimeric reads with sequence information on both the miRNA and target RNA sequence (Helwak et al., 2013). While these studies typically provided partial validation of the noncanonical sites identified within, the majority of these sites did not exhibit a functional signature upon re-analysis or extension to other data sets (Agarwal et al., 2015). These discrepancies could be caused by noncanonical sites being erroneously identified due either to systematic crosslinking biases (in which U and G nucleotides are preferentially crosslinked), or in the case of the CLASH protocols, to artificial enrichment for 3′-paired sites due to the ligation of the miRNA 3′ end to the 5′ target fragment. Another possibility is that these noncanonical sites are indeed bound with appreciable occupancy by expressed miRNAs, but for unknown reasons do not mediate repression. In any case, the results from crosslinking-based approaches, while expanding our perspective on the binding

promiscuity of some miRNAs, have not yielded a better quantitative model of miRNA-mediated repression more generally.

Understanding miRNA targeting through biochemical principles

The challenges and limitations thus far described for using in vivo data (either functional data such as RNA-seq, or in vivo binding data such as CLIP or CLASH) to understand miRNA highlight two missing pieces of information: 1) the true binding profile of any particular miRNA, and 2) the quantitative relationship between miRNA–target RNA binding and the efficacy of repression and downstream repression. Indeed, while a number of studies have implemented formal biochemical models relating target repression to miRNA concentrations, target concentrations, miRNA–target dissociation constant (K_D) values, and degradation rates (Bosson et al., 2014; Denzler et al., 2016; Jens and Rajewsky, 2014; Mukherji et al., 2011; Schmiedel et al., 2015; Wee et al., 2012), the models utilized by each make distinct sets of assumptions and estimate their parameters differently, underscoring the lack of a consistent framework for understanding miRNA targeting from a biochemical perspective. Indeed, several proposed aspects of miRNA biology have prompted controversy and debate, these being the idea that miRNAs create gene expression thresholds for their targets (Ebert and Sharp, 2012; Mukherji et al., 2011), that miRNA repression reduces the intrinsic noise of target mRNA expression (Hausser and Zavolan, 2014; Schmiedel et al., 2015), and that individual mRNA targets sites can act as competing endogenous RNAs (ceRNAs) that bind miRNAs and sequester their repression (Ala et al., 2013; Salmena et al., 2011; Tay et al., 2014). For each of these ideas, there are numerous published studies providing biochemical theory in their support (Jens and Rajewsky, 2014; Jost et al., 2013; Mukherji et al., 2011; Yuan et al., 2015). This underscores a secondary

benefit to developing an accurate framework for modeling miRNA-mediated repression beyond that of being able to accurately predict repression, as this framework itself could be tested for whether such behaviors occur.

As suggested above, construction of an informative biochemical model of miRNA-mediated repression requires measurement of binding affinities between Ago–miRNA complexes, and their target RNAs. There are numerous methods by which to measure the affinity between a protein and a nucleic acid (Jarmoskaite et al., 2020), with some of the earliest examples being nitrocellulose filter-binding (Riggs et al., 1970) and electrophoretic mobility shift assay (EMSA) (Fried and Crothers, 1981; Garner and Revzin, 1981), which were both first applied to studying the binding of the *E. coli* lac repressor to operator DNA. Two features readily distinguish the ease of applying such quantitative biochemical approaches to understanding the biology of miRNAs in comparison to the lac operon. The first is the nature of how these two regulatory modes differ—miRNA targeting necessarily involves understanding how one miRNA sequence interacts with a large diversity of RNA sequences embedded within expressed transcripts, rather than a single stretch of DNA within a small bacterial genome, which means that the number of required binding affinity measurements might be much greater. The second distinguishing feature is the added experimental difficulty of purifying a defined Ago–miRNA complex, in comparison to purifying either a transcription factor or an RBP. Indeed, when the crystal structures of yeast Ago (Nakanishi et al., 2012) and human Ago2 (Elkayam et al., 2012; Schirle and Macrae, 2012) were solved, each of the protein preparations contained a large fraction of contaminating sRNAs, coming either anomalously from the bacterial expression system (Nakanishi et al., 2012) or from the endogenous small RNA pathways from which the protein was purified (Elkayam et al., 2012; Schirle and Macrae, 2012). One group addressed this

issue by separation of the unloaded and loaded human Ago by size exclusion chromatography and incubating the unloaded population with excess single-stranded miR-20a, enabling crystallography with a biologically relevant miRNA (Elkayam et al., 2012). A clear drawback of this approach, however, was the uncertainty of whether the product of loading a single-stranded RNA into a purified Argonaute protein, absent any accessory factors used for loading in vivo, was representative of the functional, biological complex.

A clear methodological solution to the challenge of Ago–miRNA complex purification came shortly thereafter, whereby an in-lysate loading reaction was incubated with a "capture" oligo immobilized to beads, such that the Ago–miRNA complexes with a particular guide sequence could be selectively retained on the beads while other Ago–miRNA complexes (as well as the remaining constituency of the lysate) could be removed (Flores-Jasso et al., 2013). Further incubation of the beads with a "competitor" oligo with perfect complementarity to the capture oligo enabled selective elution of the Ago–miRNA complex. Subsequent removal of the competitor oligo using size-exclusion chromatography yielded a purified Ago–miRNA complex with a defined sequence, suitable for quantitative study through the application of binding, kinetic, or enzymatic assays. Indeed, the development of this technique enabled a biochemical study of both fly and mouse Ago2, each loaded with let-7a, that provided unprecedented insight into the contribution of each miRNA position to both the binding and catalysis of target slicing, and enabled a quantitative comparison of the biochemistry of the two Ago–miRNA complexes (Wee et al., 2012). In particular, the finding that the catalytic rate constant (k_{cat}) of cleavage was extremely similar to dissociation rate constant (k_{off}) for mouse Ago2–let-7a with a perfectly complementary target, but was almost 700-fold greater than k_{off} for the siRNA-loading fly Ago2

provided quantitative evidence for how these two superficially similar enzymes have been evolutionarily tuned for their respective biological pathways (Wee et al., 2012).

The capture–competitor method enabled numerous subsequent biochemical and crystallographic studies, which have together provided a more refined picture of this protein–RNA complex. In particular, studies employing single-molecule biochemistry with fluorescently tagged Ago–miRNA complexes and target RNAs have shown that miRNAs can mediate transient association through pairing only to miRNA nucleotides 2–4 (Chandradoss et al., 2015), and additionally that nucleotides 2–5 constitute a "sub-seed" that enables target binding at rates within 1–2 orders of magnitude of molecular diffusion (Jo et al., 2015; Salomon et al., 2015). These findings are consistent with published crystal structures showing, in the absence of target binding, that miRNA nucleotides 2–5 are pre-organized into a near-helical conformation (Schirle et al., 2014), and as well that nucleotides 6 and 7 exhibit significant de-stacking compared to their proceeding nucleotides (Elkayam et al., 2012).

Structural studies have also shown that the deformation of the 3′ portion of the miRNA seed prior to target pairing is caused by helix-7 of the protein (Klum et al., 2018; Schirle et al., 2014), and that its movement enables pairing to propagate through the rest of the seed region (i.e., through to nucleotide 8), thereby extending the dwell time of the target from ~0.1–1 seconds (s) to ~5–250 s (Chandradoss et al., 2015; Salomon et al., 2015). They have additionally shed light on the nature of the preference for an A nucleotide across from position 1 of the miRNA irrespective of its nucleotide identify, identifying a binding pocket formed through the interface of the MID and PIWI domains within which an ordered array of water molecules specifically recognize the adenosine N6 amine (Schirle et al., 2015). Finally, crystallography studies comparing the structures of Ago–miRNA complexes bound to targets with iteratively

more complementarity beyond nucleotide 8 have provided a physical basis for understanding why pairing to the central region of the miRNA contributes so little to miRNA targeting: nucleotides 9–11 are conformationally excluded by a central gate, with solvent exposure returning at nucleotides 13–16 upon seed binding (Schirle et al., 2014; Sheu-Gruttadauria et al., 2019a). In addition, these structural data have partially informed a proposal that target cleavage might progress through initial seed pairing followed by a secondary nucleation of pairing within the 3′ end, with back-propagation of the secondary helix causing opening of the central gate and allowing access to the phosphodiester linkage bridging nucleotides 10 and 11 (Bartel, 2018).

The studies thus described serve to illustrate the myriad ways in which the conformational and binding properties of a miRNA are fundamentally changed upon loading into an Argonaute protein. Indeed, this remodeling by Ago provides a clear rationale for why models of miRNA effects based on predicted pairing stability (Khorshid et al., 2013; Rajewsky and Socci, 2004) have not been as successful as those that evaluate pairing to particular positions of the miRNA (Agarwal et al., 2015; Lewis et al., 2003). However, those biochemical studies which performed biochemical assays with more than one miRNA sequence demonstrated clear differences in the k_{on} and k_{off} for seed pairing between let-7a and miR-21 (Salomon et al., 2015), in the k_{off} for seed pairing between miR-27 and both let-7a and miR-122 (Sheu-Gruttadauria et al., 2019b), in the propensity for cleavage (given by the ratio k_{cat}/k_{off}) between let-7a and let-7b (Jo et al., 2015), and in the propensity for differential 3′ pairing between different sequence variants of miR-122 (Sheu-Gruttadauria et al., 2019a). These results, when considered together with the finding that predicted seed-pairing stability (SPS) is a useful-but-imperfect correlate of in vivo miRNA repression between different miRNAs (Garcia et al., 2011), support a model in which the primary reason why miRNA target prediction remains poor is a lack of understanding

of how the predicted energetics of a guide sequence are transformed once in complex with Ago, the extent to which this transformation differs between miRNAs, and which miRNA sequence features are responsible for any such differences.

Building such an understanding would require many more measurements than those present in the studies thus discussed. To this end, a more recent study performing high-throughput biochemistry using a modified Illumina sequencing platform measured ~20,000 K_D, k_{on}, and slicing k_{cat} values for let-7a and miR-21, with ~2,000 and ~5,000 target sites, respectively, drawn from the top-predicted targets with several miRNA target prediction algorithms (Becker et al., 2019). While these data provide richer quantitative insights into the differences in binding between let-7a and miR-21, and would therefore be expected to improve prediction of the efficacy of both miRNA-mediated repression and target slicing for both these miRNAs, the predetermined nature of the pool of target RNAs queried for both miRNAs means that some functional site types or relevant sequence features might be missed in these experiments, simply due to their lack of representation within the target pool. To this end, an experimental technique enabling assessment of a vast number of putative target sites might provide a means to improve miRNA target prediction.

Random sequence–based, high-throughput biochemistry

The variety of contemporary methods for sequence-motif discovery find their conceptual origin in a technique developed 30 years ago called selective evolution of ligands by exponential enrichment (SELEX). The method utilizes a population of partially or fully randomized RNA molecules that are iteratively subjected to rounds of binding-based selection, reverse transcription, amplification, and in vitro transcription, thereby enriching for those few RNA

molecules in the initial pool with the greatest binding affinity for the desired binding partner (Ellington and Szostak, 1990; Tuerk and Gold, 1990). In its early form, the results of the experiment could only be queried by Sanger sequencing of either the final pool (Blackwell and Weintraub, 1990) or a handful of molecules cloned from the final pool (Fields et al., 1997; Jin et al., 2003), such that only qualitative information could be derived regarding the preferred binding sequence of a given protein.

With the advent of high-throughput sequencing, a variety of related methods were developed such as high-throughput SELEX (HT-SELEX) (Zhao et al., 2009), Bind-n-Seq (Zykovich et al., 2009), and SELEX-seq (Slattery et al., 2011), in which a dsDNA pool is sequenced in its initial state and after each round of selection for transcription factor (TF) binding. These approaches enabled a richer and more quantitative approach for learning TF-binding specificity, albeit with some drawbacks: because the early-round pools tended to contain a large fraction of non-specific binding, and because the later rounds were mostly dominated by the highest-affinity sequences, the medium-to-low-affinity sequences would either be missed or inaccurately quantified. Indeed, a recent computational analysis pipeline employed for analyzing single-round HT-SELEX data has been able to quantify relative K_D values for individual transcription factors within a 160-fold range, indicating that with sophisticated biophysical modeling and statistical treatment, apparent limitations of the assay can be overcome (Rastogi et al., 2018).

The high-throughput, single-round SELEX approach was subsequently applied for the purposes of studying RBPs using a pool of RNA molecules with 20 or 40 random nucleotide positions (Dominguez et al., 2018; Lambert et al., 2014). The technique was named RNA Bind-n-Seq (RBNS), because, like Bind-n-Seq, the protocol included multiple binding reactions per

RBP studied, over which the RBP concentration was varied to obtain quantitative enrichments of motifs at different levels of saturation of the library (Lambert et al., 2014). In addition to describing the sequence preferences of RBFOX2, CELF1, and MBNL1, and showing the superiority of the RBNS-generated profiles in comparison to CLIP for predicting alternative splicing in vivo, the pioneering RBNS study identified that the *k*-mer enrichment values generated within an RBNS reaction exhibit unimodal enrichment values at intermediate RBP concentrations, owing to nonspecific binding at low RBP concentrations and to RNA library saturation at higher RBP concentrations. Indeed, the waning enrichments were shown to be qualitatively consistent with a biochemical model of the experiment (Lambert et al., 2014), suggesting that, if applied to Ago–miRNA complexes, RBNS might enable novel site discovery as well as the construction of site-type affinity profiles with accurate relative K_D values spanning the full dynamic range of binding, thereby enabling an unprecedented view into the targeting preferences of individual miRNAs.

Organization of thesis

The remaining chapters of this thesis will describe my experimental and computational work to adapt RBNS for use with human AGO–miRNA complexes, and the improvements made to our quantitative understanding of miRNA targeting as a result of these measurements. Chapter 2 describes the development of AGO-RBNS and, through collaboration with Kathy S. Lin, the generation of a biochemically informed model of miRNA targeting that outperforms all other current target prediction algorithms, as well as the construction of a convolutional neural network (CNN) that predicts relative K_D values for a miRNA of any sequence. Chapter 3 describes work performed in collaboration with Namita Bisaria to perform AGO-RBNS

experiments reporting on the contribution of the miRNA 3′ end to binding affinity, analysis of which demonstrated the existence of two distinct binding modes enabling productive 3′ pairing, and that miRNA G nucleotides shape the 3′-pairing preferences of individual miRNAs. Chapter 4 synthesizes the results spanning these chapters, and attempts to provide perspective on how further advances in understanding miRNA targeting might be achieved. The appendices serve to collect the research papers to which I have contributed in supportive roles, being A) in vitro biochemistry to confirm insights related to the specificity of miRNA biogenesis, B) modeling in service of verifying the unimodal enrichment patterns generated by RBNS experiments, C) modeling to confirm the relative dynamics of translational repression and mRNA destabilization during miRNA-mediated repression, D) modeling to demonstrate the non-physiological conditions in which ceRNAs could plausibly titrate the function of a miRNA, and E) assistance in formulating a mathematical framework describing the dynamics of poly(A) tail–length changes during the life of a eukaryotic mRNA.

References

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, e05005.

Ahmed, K., LaPierre, M.P., Gasser, E., Denzler, R., Yang, Y., Rülicke, T., Kero, J., Latreille, M., and Stoffel, M. (2017). Loss of microRNA-7a2 induces hypogonadotropic hypogonadism and infertility. J. Clin. Invest. *127*, 1061–1074.

Ala, U., Karreth, F.A., Bosia, C., Pagnani, A., Taulli, R., Léopold, V., Tay, Y., Provero, P., Zecchina, R., and Pandolfi, P.P. (2013). Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. Proc. Natl. Acad. Sci. USA *110*, 7154–7159.

Alvarez-Saavedra, E., and Horvitz, H.R. (2010). Many families of *C. elegans* microRNAs are not essential for development or viability. Curr. Biol. *20*, 367–373.

Ambros, V. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*. Cell *57*, 49–57.

Anders, G., Mackowiak, S.D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res. *40*, D180–D186.

Andolina, D., Segni, M.D., Bisicchia, E., D'Alessandro, F., Cestari, V., Ventura, A., Concepcion, C., Puglisi-Allegra, S., and Ventura, R. (2016). Effects of lack of microRNA-34 on the neural circuitry underlying the stress response and anxiety. Neuropharmacology *107*, 305–316.

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71.

Bartel, D.P. (2009). MicroRNAs: Target recognition and regulatory functions. Cell *136*, 215–233.

Bartel, D.P. (2018). Metazoan microRNAs. Cell *173*, 20–51.

Bartel, D.P., Sheng, M., Lau, L.F., and Greenberg, M.E. (1989). Growth factors and membrane depolarization activate distinct programs of early response gene expression: dissociation of *fos* and *jun* induction. Genes Dev. *3*, 304–313.

Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science *336*, 233–237.

Becker, W.R., Ober-Reynolds, B., Jouravleva, K., Jolly, S.M., Zamore, P.D., and Greenleaf, W.J. (2019). High-throughput analysis reveals rules for target RNA binding and cleavage by AGO2. Mol. Cell *75*, 741–755.e11.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. *20*, 1885–1898.

Blackwell, T., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104–1110.

Bosson, A.D., Zamudio, J.R., and Sharp, P.A. (2014). Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. Mol. Cell *56*, 347–359.

Brancati, G., and Großhans, H. (2018). An interplay of miRNA abundance and target site architecture determines miRNA activity and specificity. Nucleic Acids Res. *46*, gky201-.

Braun, J.E., Huntzinger, E., Fauser, M., and Izaurralde, E. (2011). GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. Mol. Cell *44*, 120–133.

Braun, J.E., Huntzinger, E., and Izaurralde, E. (2012). A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. Cold Spring Harb. Perspect. Biol. *4*, a012328.

Braun, J.E., Huntzinger, E., and Izaurralde, E. (2013). Ten years of progress in GW/P body research. Adv. Exp. Med. Biol. *768*, 147–163.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA–target recognition. PLOS Biol. *3*, e85.

Cao, D., and Parker, R. (2001). Computational modeling of eukaryotic mRNA turnover. RNA *7*, 1192–1212.

Chandradoss, S.D., Schirle, N.T., Szczepaniak, M., Macrae, I.J., and Joo, C. (2015). A dynamic search process underlies microRNA targeting. Cell *162*, 96–107.

Cheloufi, S., Santos, C.O.D., Chong, M.M.W., and Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584–589.

Chen, C.A., and Shyu, A. (2011). Mechanisms of deadenylation-dependent decay. Wiley Interdiscip. Rev. RNA *2*, 167–183.

Chen, G.R., Sive, H., and Bartel, D.P. (2017). A seed mismatch enhances Argonaute2-catalyzed cleavage and partially rescues severely impaired cleavage found in fish. Mol. Cell *68*, 1095–1107.e5.

Chen, Y., Boland, A., Kuzuoğlu-Öztürk, D., Bawankar, P., Loh, B., Chang, C.-T., Weichenrieder, O., and Izaurralde, E. (2014). A DDX6-CNOT1 complex and W-binding pockets in CNOT9 reveal direct links between miRNA target recognition and silencing. Mol. Cell *54*, 737–750.

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

Chi, S.W., Hannon, G.J., and Darnell, R.B. (2012). An alternative mode of microRNA target recognition. Nat. Struct. Mol. Biol. *19*, 321–327.

Chowdhury, A., Mukhopadhyay, J., and Tharun, S. (2007). The decapping activator Lsm1p-7p–Pat1p complex has the intrinsic ability to distinguish between oligoadenylated and polyadenylated RNAs. RNA *13*, 998–1016.

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., et al. (2010). A novel miRNA processing pathway independent of dicer requires Argonaute2 catalytic activity. Science *328*, 1694–1698.

Connelly, S., and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev. *2*, 440–452.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Davis, E., Caiment, F., Tordoir, X., Cavaillé, J., Ferguson-Smith, A., Cockett, N., Georges, M., and Charlier, C. (2005). RNAi-mediated allelic trans-interaction at the imprinted *Rtl1*/*Peg11* locus. Curr. Biol. *15*, 743–749.

Denzler, R., McGeary, S.E., Title, A.C., Agarwal, V., Bartel, D.P., and Stoffel, M. (2016). Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. Mol. Cell *64*, 565–579.

Dexheimer, P.J., Wang, J., and Cochella, L. (2020). Two microRNAs are sufficient for embryonic patterning in *C. elegans*. Curr. Biol. *30*, 1–8.

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Nostrand, E.L.V., Pratt, G.A., et al. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Mol. Cell *70*, 854–867.e9.

Dooley, J., Garcia-Perez, J.E., Sreenivasan, J., Schlenner, S.M., Vangoitsenhoven, R., Papadopoulou, A.S., Tian, L., Schonefeldt, S., Serneels, L., Deroose, C., et al. (2015). The microRNA-29 family dictates the balance between homeostatic and pathological glucose handling in diabetes and obesity. Diabetes *65*, 53–61.

Ebert, M.S., and Sharp, P.A. (2012). Roles for microRNAs in conferring robustness to biological processes. Cell *149*, 515–524.

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

Eichhorn, S.W., Subtelny, A.O., Kronja, I., Kwasnieski, J.C., Orr-Weaver, T.L., and Bartel, D.P. (2016). mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. eLife *5*, e16955.

Eisen, T.J., Eichhorn, S.W., Subtelny, A.O., Lin, K.S., McGeary, S.E., Gupta, S., and Bartel, D.P. (2020a). The dynamics of cytoplasmic mRNA metabolism. Mol. Cell *77*, 786–799.e10.

Eisen, T.J., Eichhorn, S.W., Subtelny, A.O., and Bartel, D.P. (2020b). MicroRNAs cause accelerated decay of short-tailed target mRNAs. Mol. Cell *77*, 775–785.e8.

Elbashir, S.M., Lendeckel, W., and Tuschl, T. (2001a). RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. *15*, 188–200.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001b). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494–498.

Elkayam, E., Kuhn, C.-D., Tocilj, A., Haase, A.D., Greene, E.M., Hannon, G.J., and Joshua-Tor, L. (2012). The structure of human Argonaute-2 in complex with miR-20a. Cell *150*, 100–110.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.

Eulalio, A., Huntzinger, E., and Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. Nat. Struct. Mol. Biol. *15*, 346–353.

Farh, K.K.-H., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science *310*, 1817–1821.

Fields, D.S., He, Y., Al-Uzri, A.Y., and Stormo, G.D. (1997). Quantitative specificity of the Mnt repressor. J. Mol. Biol. *271*, 178–194.

Fire, A., Albertson, D., Harrison, S.W., and Moerman, D.G. (1991). Production of antisense RNA leads to effective and specific inhibition of gene expression in *C. elegans* muscle. Development *113*, 503–514.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2013). Rapid and specific purification of Argonaute-small RNA complexes from crude cell lysates. RNA *19*, 271–279.

Fried, M., and Crothers, D.M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res. *9*, 6505–6525.

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

Gaidatzis, D., Nimwegen, E. van, Hausser, J., and Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinforma. *8*, 69.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs. Nat. Struct. Mol. Biol. *18*, 1139–1146.

Garner, M.M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res. *9*, 3047–3060.

Gasperini, M., Tome, J.M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet. *21*, 292–310.

Giraldez, A.J. (2006). Zebrafish miR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75–79.

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

Goldstrohm, A.C., and Wickens, M. (2008). Multifunctional deadenylase complexes diversify mRNA control. Nat. Rev. Mol. Cell Biol. *9*, 337–344.

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

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. Mol. Cell *54*, 1042–1054.

Gumienny, R., and Zavolan, M. (2015). Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. Nucleic Acids Res. *43*, 1380–1391.

Guo, S., and Kemphues, K.J. (1995). *par-1*, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. Cell *81*, 611–620.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Guo, Y., Liu, J., Elfenbein, S.J., Ma, Y., Zhong, M., Qiu, C., Ding, Y., and Lu, J. (2015). Characterization of the mammalian miRNA turnover landscape. Nucleic Acids Res. *43*, 2326–2341.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Haley, B., and Zamore, P.D. (2004). Kinetic analysis of the RNAi enzyme complex. Nat. Struct. Mol. Biol. *11*, 599–606.

Han, Y.-C., Vidigal, J.A., Mu, P., Yao, E., Singh, I., González, A.J., Concepcion, C.P., Bonetti, C., Ogrodowski, P., Carver, B., et al. (2015). An allelic series of miR-17∼92–mutant mice uncovers functional specialization and cooperation among members of a microRNA polycistron. Nat. Genet. *47*, 766–775.

Hasuwa, H., Ueda, J., Ikawa, M., and Okabe, M. (2013). miR-200b and miR-429 function in mouse ovulation and are essential for female fertility. Science *341*, 71–73.

Hausser, J., and Zavolan, M. (2014). Identification and consequences of miRNA–target interactions—beyond repression of gene expression. Nat. Rev. Genet. *15*, 599–612.

Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., and Zavolan, M. (2009). Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C–miRNA complexes and the degradation of miRNA targets. Genome Res. *19*, 2009–2020.

Heidersbach, A., Saxby, C., Carver-Moore, K., Huang, Y., Ang, Y.-S., Jong, P.J. de, Ivey, K.N., and Srivastava, D. (2013). microRNA-1 regulates sarcomere formation and suppresses smooth muscle gene expression in the mammalian heart. eLife *2*, e01323.

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

Hentze, M.W. (1991). Determinants and regulation of cytoplasmic mRNA stability in eukaryotic cells. Biochim. Biophys. Acta *1090*, 281–292.

Hentze, M., Caughman, S., Rouault, T., Barriocanal, J., Dancis, A., Harford, J., and Klausner, R. (1987). Identification of the iron-responsive element for the translational regulation of human ferritin mRNA. Science *238*, 1570–1573.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. *36*, 39–56.

Huberdeau, M.Q., Zeitler, D.M., Hauptmann, J., Bruckmann, A., Fressigné, L., Danner, J., Piquet, S., Strieder, N., Engelmann, J.C., Jannot, G., et al. (2017). Phosphorylation of Argonaute proteins affects mRNA binding and is essential for microRNA-guided gene silencing in vivo. EMBO J. *36*, 2088–2106.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174–1189.

Hutvágner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056–2060.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Jan, C.H., Friedman, R.C., Ruby, J.G., and Bartel, D.P. (2011). Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. Nature *469*, 97–101.

Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P.P., and Herschlag, D. (2020). How to measure and evaluate binding affinities. eLife *9*, e57264.

Jee, D., Yang, J.-S., Park, S.-M., Farmer, D.T., Wen, J., Chou, T., Chow, A., McManus, M.T., Kharas, M.G., and Lai, E.C. (2018). Dual strategies for Argonaute2-mediated biogenesis of erythroid miRNAs underlie conserved requirements for slicing in mammals. Mol. Cell *69*, 265–278.e6.

Jens, M., and Rajewsky, N. (2014). Competition between target sites of regulators shapes post-transcriptional gene regulation. Nat. Rev. Genet. *16*, 113–126.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. EMBO J. *22*, 905–912.

Jo, M.H., Shin, S., Jung, S.-R., Kim, E., Song, J.-J., and Hohng, S. (2015). Human Argonaute 2 has diverse reaction pathways on target RNAs. Mol. Cell *59*, 117–124.

Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nat. Rev. Genet. *16*, 421–433.

Jost, D., Nowojewski, A., and Levine, E. (2013). Regulating the many to benefit the few: role of weak small RNA targets. Biophys. J. *104*, 1773–1782.

Khorshid, M., Hausser, J., Zavolan, M., and Nimwegen, E. van (2013). A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nat. Methods *10*, 253–255.

Kim, D., Sung, Y.M., Park, J., Kim, S., Kim, J., Park, J., Ha, H., Bae, J.Y., Kim, S., and Baek, D. (2016). General rules for functional microRNA targeting. Nat. Genet. *48*, 1517–1526.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kingston, E.R., and Bartel, D.P. (2019). Global analyses of the dynamics of mammalian microRNA metabolism. Genome Res. *29*, 1777–1790.

Klum, S.M., Chandradoss, S.D., Schirle, N.T., Joo, C., and MacRae, I.J. (2018). Helix-7 in Argonaute2 shapes the microRNA seed region for rapid target recognition. EMBO J. *37*, 75–88.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47*, gky1141-.

Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., Piedade, I. da, Gunsalus, K.C., Stoffel, M., et al. (2005). Combinatorial microRNA target predictions. Nat. Genet. *37*, 495–500.

Kretov, D.A., Walawalkar, I.A., Mora-Martin, A., Shafik, A.M., Moxon, S., and Cifuentes, D. (2020). Ago2-dependent processing allows miR-451 to evade the global microRNA turnover elicited during erythropoiesis. Mol. Cell *78*, 317–328.e6.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. Proc. Natl. Acad. Sci. USA *108*, 10010–10015.

Kuhn, S., Johnson, S.L., Furness, D.N., Chen, J., Ingham, N., Hilton, J.M., Steffes, G., Lewis, M.A., Zampini, V., Hackney, C.M., et al. (2011). miR-96 regulates the progression of differentiation in mammalian cochlear inner and outer hair cells. Proc. Natl. Acad. Sci. USA *108*, 2355–2360.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853–858.

Lai, E.C. (2002). MicroRNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat. Genet. *30*, 363–364.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–900.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858–862.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science *294*, 862–864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell *75*, 843–854.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

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

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

Lewis, M.A., Quint, E., Glazier, A.M., Fuchs, H., Angelis, M.H.D., Langford, C., Dongen, S. van, Abreu-Goodger, C., Piipari, M., Redshaw, N., et al. (2009). An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. Nat. Genet. *41*, 614–618.

Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ *2*, e270.

Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L., and Betel, D. (2011). Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev. *25*, 2173–2186.

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

Liu, N., Bezprozvannaya, S., Williams, A.H., Qi, X., Richardson, J.A., Bassel-Duby, R., and Olson, E.N. (2008). microRNA-133a regulates cardiomyocyte proliferation and suppresses smooth muscle gene expression in the heart. Genes Dev. *22*, 3242–3254.

Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., and Rudensky, A.Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol. Cell *48*, 760–770.

Logan, J., Falck-Pedersen, E., Darnell, J.E., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc. Natl. Acad. Sci. USA *84*, 8306–8310.

Madison, B.B., Liu, Q., Zhong, X., Hahn, C.M., Lin, N., Emmett, M.J., Stanger, B.Z., Lee, J.-S., and Rustgi, A.K. (2013). LIN28B promotes growth and tumorigenesis of the intestinal epithelium via Let-7. Genes Dev. *27*, 2233–2245.

Majoros, W.H., and Ohler, U. (2007). Spatial preferences of microRNA targets in 3′ untranslated regions. BMC Genom. *8*, 152.

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

Mencía, Á., Modamio-Høybjør, S., Redshaw, N., Morín, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L.A., Castillo, I. del, Steel, K.P., Dalmay, T., et al. (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat. Genet. *41*, 609–613.

Meyer, S., Temme, C., and Wahle, E. (2010). Messenger RNA turnover in eukaryotes: pathways and enzymes. Crit. Rev. Biochem. Mol. Biol. *39*, 197–216.

Miska, E.A., Alvarez-Saavedra, E., Abbott, A.L., Lau, N.C., Hellman, A.B., McGonagle, S.M., Bartel, D.P., Ambros, V.R., and Horvitz, H.R. (2007). Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. PLoS Genet. *3*, e215.

Moran, Y., Agron, M., Praher, D., and Technau, U. (2017). The evolutionary origin of plant and animal microRNAs. Nat. Ecol. Evol. *1*, 0027.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the *lin-4* RNA. Cell *88*, 637–646.

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

Nakanishi, K., Weinberg, D.E., Bartel, D.P., and Patel, D.J. (2012). Structure of yeast Argonaute with guide RNA. Nature *486*, 368–374.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. Plant Cell *2*, 279.

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

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Müller, P., et al. (2000). Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. Nature *408*, 86–89.

Penzkofer, D., Bonauer, A., Fischer, A., Tups, A., Brandes, R.P., Zeiher, A.M., and Dimmeler, S. (2014). Phenotypic characterization of miR-92a−/− mice reveals an important function of miR-92a in skeletal development. PLOS ONE *9*, e101153.

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F.P., Chang, Y.-C., Madugundu, A.K., Pandey, A., and Salzberg, S.L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. *19*, 208.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Proudfoot, N.J. (1989). How RNA polymerase II terminates transcription in higher eukaryotes. Trends Biochem. Sci. *14*, 105–110.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501–512.

Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Dev. Biol. *267*, 529–535.

Rastogi, C., Rube, H.T., Kribelbauer, J.F., Crocker, J., Loker, R.E., Martini, G.D., Laptenko, O., Freed-Pastor, W.A., Prives, C., Stern, D.L., et al. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. Proc. Natl. Acad. Sci. USA *115*, 201714376.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901–906.

Riggs, A.D., Suzuki, H., and Bourgeois, S. (1970). lac repressor-operator interaction I. Equilibrium studies. J. Mol. Biol. *48*, 67–83.

Rivas, F.V., Tolia, N.H., Song, J.-J., Aragon, J.P., Liu, J., Hannon, G.J., and Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. Nat. Struct. Mol. Biol. *12*, 340–349.

Rooij, E. van, Sutherland, L.B., Qi, X., Richardson, J.A., Hill, J., and Olson, E.N. (2007). Control of stress-dependent cardiac growth and gene expression by a microRNA. Science *316*, 575–579.

Rosenthal, E.T., Tansey, T.R., Ruderman, J.V., and Gottesman, M. (1983). Sequence-specific adenylations and deadenylations accompany changes in the translation of maternal messenger RNA after fertilization of *Spisula* oocytes. J. Mol. Biol. *166*, 309–327.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Salomon, W.E., Jolly, S.M., Moore, M.J., Zamore, P.D., and Serebrov, V. (2015). Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. Cell *162*, 84–95.

Sanuki, R., Onishi, A., Koike, C., Muramatsu, R., Watanabe, S., Muranishi, Y., Irie, S., Uneo, S., Koyasu, T., Matsui, R., et al. (2011). miR-124a is required for hippocampal axogenesis and retinal cone survival through Lhx2 suppression. Nat. Neurosci. *14*, 1125–1134.

Schirle, N.T., and Macrae, I.J. (2012). The crystal structure of human Argonaute2. Science *336*, 1037–1040.

Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. Science *346*, 608–613.

Schirle, N.T., Sheu-Gruttadauria, J., Chandradoss, S.D., Joo, C., and MacRae, I.J. (2015). Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. eLife *4*, e07646.

Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D.S., and Oudenaarden, A. van (2015). MicroRNA control of protein expression noise. Science *348*, 128–132.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Sheng, M., and Greenberg, M.E. (1990). The regulation and function of c-*fos* and other immediate early genes in the nervous system. Neuron *4*, 477–485.

Sheu-Gruttadauria, J., Xiao, Y., Gebert, L.F., and MacRae, I.J. (2019a). Beyond the seed: structural basis for supplementary microRNA targeting by human Argonaute2. EMBO J. *38*, e101153.

Sheu-Gruttadauria, J., Pawlica, P., Klum, S.M., Wang, S., Yario, T.A., Oakdale, N.T.S., Steitz, J.A., and MacRae, I.J. (2019b). Structural basis for target-directed microRNA degradation. Mol. Cell *75*, 1243–1255.e7.

Shibata, M., Nakao, H., Kiyonari, H., Abe, T., and Aizawa, S. (2011). MicroRNA-9 regulates neurogenesis in mouse telencephalon by targeting multiple transcription factors. J. Neurosci. *31*, 3407–3422.

Shin, C., Nam, J.-W., Farh, K.K.-H., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. Mol. Cell *38*, 789–802.

Siegel, J.J., and Amon, A. (2012). New insights into the troubles of aneuploidy. Annu. Rev. Cell Dev. Biol. *28*, 189–214.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Smith, K.M., Guerau-de-Arellano, M., Costinean, S., Williams, J.L., Bottoni, A., Cox, G.M., Satoskar, A.R., Croce, C.M., Racke, M.K., Lovett-Racke, A.E., et al. (2012). miR-29ab1 deficiency identifies a negative feedback loop controlling Th1 bias that is dysregulated in multiple sclerosis. J. Immunol. *189*, 1567–1576.

Sonenberg, N., Morgan, M.A., Merrick, W.C., and Shatkin, A.J. (1978). A polypeptide in eukaryotic initiation factors that crosslinks specifically to the 5′-terminal cap in mRNA. Proc. Natl. Acad. Sci. USA *75*, 4843–4847.

Song, R., Walentek, P., Sponer, N., Klimke, A., Lee, J.S., Dixon, G., Harland, R., Wan, Y., Lishko, P., Lize, M., et al. (2014). miR-34/449 miRNAs are required for motile ciliogenesis by repressing *cp110*. Nature *510*, 115–120.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell *9*, 3273–3297.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell *123*, 1133–1146.

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

Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E.V., Patel, D.J., and Oost, J. van der (2014). The evolutionary journey of Argonaute proteins. Nat. Struct. Mol. Biol. *21*, 743–753.

Tafer, H., Ameres, S.L., Obernosterer, G., Gebeshuber, C.A., Schroeder, R., Martinez, J., and Hofacker, I.L. (2008). The impact of target site accessibility on the design of effective siRNAs. Nat. Biotechnol. *26*, 578–583.

Tay, Y., Rinn, J., and Pandolfi, P.P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature *505*, 344–352.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505–510.

Tuschl, T., Zamore, P.D., Lehmann, R., Bartel, D.P., and Sharp, P.A. (1999). Targeted mRNA degradation by double-stranded RNA in vitro. Genes Dev. *13*, 3191–3197.

Ui-Tei, K., Naito, Y., Nishi, K., Juni, A., and Saigo, K. (2008). Thermodynamic stability and Watson–Crick base pairing in the seed duplex are major determinants of the efficiency of the siRNA-based off-target effect. Nucleic Acids Res. *36*, 7100–7109.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706–709.

Wang, H., Moyano, A.L., Ma, Z., Deng, Y., Lin, Y., Zhao, C., Zhang, L., Jiang, M., He, X., Ma, Z., et al. (2017). miR-219 cooperates with miR-338 in myelination and promotes myelin repair in the CNS. Dev. Cell *40*, 566–582.e5.

Watson, J.D., and Crick, F.H.C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. Cell *151*, 1055–1067.

Wei, Y., Peng, S., Wu, M., Sachidanandam, R., Tu, Z., Zhang, S., Falce, C., Sobie, E.A., Lebeche, D., and Zhao, Y. (2014). Multifaceted roles of miR-1s in repressing the fetal gene program in the heart. Cell Res. *24*, 278–292.

Weill, L., Belloc, E., Bava, F.-A., and Méndez, R. (2012). Translational control by changes in poly(A) tail length: recycling mRNAs. Nat. Struct. Mol. Biol. *19*, 577–585.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in C. elegans. Cell *75*, 855–862.

Wu, L., Nguyen, L.H., Zhou, K., Soysa, T.Y. de, Li, L., Miller, J.B., Tian, J., Locker, J., Zhang, S., Shinoda, G., et al. (2015). Precise *let-7* expression levels balance organ regeneration against tumor suppression. eLife *4*, e09431.

Yekta, S., Shih, I., and Bartel, D.P. (2004). MicroRNA-directed cleavage of *HOXB8* mRNA. Science *304*, 594–596.

Yuan, Y., Liu, B., Xie, P., Zhang, M.Q., Li, Y., Xie, Z., and Wang, X. (2015). Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. Proc. Natl. Acad. Sci. USA *112*, 3158–3163.

Zamenhof, S., Brawerman, G., and Chargaff, E. (1952). On the desoxypentose nucleic acids from several microorganisms. Biochim. Biophys. Acta *9*, 402–405.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25–33.

Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLOS Comput. Biol. *5*, e1000590.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. Nucleic Acids Res. *37*, e151–e151.

Chapter 2.

The biochemical basis of microRNA targeting efficacy

Sean E. McGeary[1,2,3]*, Kathy S. Lin[1,2,3,4]*, Charlie Y. Shi[1,2,3], Thy M. Pham[1,2,3], Namita Bisaria[1,2,3], Gina M. Kelley[1,2,3], and David P. Bartel[1,2,3,4]

[1]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*These authors contributed equally to this work.

S.E.M. developed AGO-RBNS and associated analyses, which he implemented with help from T.P. and N.B. K.S.L. devised and implemented the biochemical model and CNN. C.Y.S., G.M.K., and T.P. performed transfection and sequencing experiments. C.Y.S. and S.E.M. designed and performed the massively parallel reporter assay. S.E.M., K.S.L., and D.P.B. designed the study and wrote the manuscript with input from other authors.

Published as:
McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., and Bartel, D.P. (2019). The biochemical basis of microRNA targeting efficacy. Science *366*, eaav1741.

Abstract

MicroRNAs (miRNAs) act within Argonaute proteins to guide repression of messenger RNA targets. Although various approaches have provided insight into target recognition, the sparsity of miRNA–target affinity measurements has limited understanding and prediction of targeting efficacy. Here, we adapted RNA bind-n-seq to enable measurement of relative binding affinities between Argonaute–miRNA complexes and all sequences ≤12 nucleotides in length. This approach revealed noncanonical target sites specific to each miRNA, miRNA-specific differences in canonical target-site affinities, and a 100-fold impact of dinucleotides flanking each site. These data enabled construction of a biochemical model of miRNA-mediated repression, which was extended to all miRNA sequences using a convolutional neural network. This model substantially improved prediction of cellular repression, thereby providing a biochemical basis for quantitatively integrating miRNAs into gene-regulatory networks.

Introduction

MicroRNAs (miRNAs) are ~22-nucleotide (nt) regulatory RNAs that derive from hairpin regions of precursor transcripts (Bartel, 2018). Each miRNA associates with an Argonaute (AGO) protein to form a silencing complex, in which the miRNA pairs to sites within target transcripts and the AGO protein promotes destabilization and/or translational repression of bound transcripts (Jonas and Izaurralde, 2015). miRNAs are grouped into families on the basis of the sequence of their extended seed (nucleotides 2–8 of the miRNA), which is the region of the miRNA most important for target recognition (Bartel, 2009). The 90 most broadly conserved miRNA families of mammals each have an average of >400 preferentially conserved targets, such that mRNAs from most human genes are conserved targets of at least one miRNA

(Friedman et al., 2009). Most of these 90 broadly conserved families are required for normal development or physiology, as shown by knockout studies in mice (Bartel, 2018).

Deeper understanding of these numerous biological functions would be facilitated by a better understanding of miRNA targeting efficacy, with the ultimate goal of correctly predicting the effects of each miRNA on the output of each expressed gene. In principle, targeting efficacy should be a function of the affinity between AGO–miRNA complexes and their target sites, in that greater affinity to a target site would cause increased occupancy at that site and thus increased repression of the target mRNA. Until very recently, binding affinities have been known for only a few target sequences of only three miRNAs (Chandradoss et al., 2015; Jo et al., 2015; Klum et al., 2018; Salomon et al., 2015; Schirle et al., 2014, 2015; Wee et al., 2012). In a recent study, high-throughput imaging and cleavage analyses provide extensive binding and slicing data for two of these three miRNAs, let-7a and miR-21 (Becker et al., 2019). Although these measurements provide insight and enable a quantitative model that predicts the efficiency of miR-21–directed slicing in cells (Becker et al., 2019), the sparsity of binding-affinity data still limits insight into how targeting might differ between different miRNAs and prevents construction of an informative biochemical model of targeting efficacy relevant to the vastly more prevalent, non-slicing mode of miRNA-mediated repression.

With insufficient affinity measurements, the most informative models of targeting efficacy rely instead on indirect, correlative approaches. These models focus on mRNAs with canonical 6–8-nt sites matching the miRNA seed region (Figure 1A) and train on features known to correlate with targeting efficacy (including the type of site as well as various features of site context, mRNAs, and miRNAs), by using datasets that monitor mRNA changes that occur after introducing a miRNA (Agarwal et al., 2015; Grimson et al., 2007; Gumienny and Zavolan, 2015; Paraskevopoulou et al., 2013). Although the correlative model implemented in TargetScan7

performs as well as the best in vivo cross-linking approaches at predicting mRNAs most responsive to miRNA perturbation, it nonetheless explains only a small fraction of the mRNA changes observed upon introducing a miRNA [coefficient of determination (r^2) = 0.14] (Agarwal et al., 2015). This low value indicates that prediction of targeting efficacy has room for improvement, even when accounting for the fact that experimental noise and secondary effects of inhibiting direct targets place a ceiling on the variability attributable to direct targeting. Therefore, we adapted RNA bind-n-seq (RBNS) (Lambert et al., 2014) and a convolutional neural network (CNN) to the study of miRNA–target interactions, with the goal of obtaining the quantity and diversity of affinity measurements needed to better understand and predict miRNA targeting efficacy.

The site-affinity profile of miR-1

As previously implemented, RBNS provides qualitative relative binding measurements for an RNA-binding protein to a virtually exhaustive list of binding sites (Dominguez et al., 2018; Lambert et al., 2014). A purified RNA-binding protein is incubated with a large library of RNA molecules that each contain a central random-sequence region flanked by constant primer-binding regions. After reaching binding equilibrium, the protein is pulled down and any co-purifying RNA molecules are reverse transcribed, amplified, and sequenced. To extend RBNS to AGO–miRNA complexes (Figure 1B), we purified human AGO2 loaded with miR-1 (Flores-Jasso et al., 2013) (Figure S1A) and set up five binding reactions, each with a different concentration of AGO2–miR-1 (range, 7.3–730 pM, logarithmically spaced) and a constant concentration of an RNA library with a 37-nt random-sequence region (100 nM). We also modified the protein-isolation step of the RBNS protocol, replacing protein pull down with

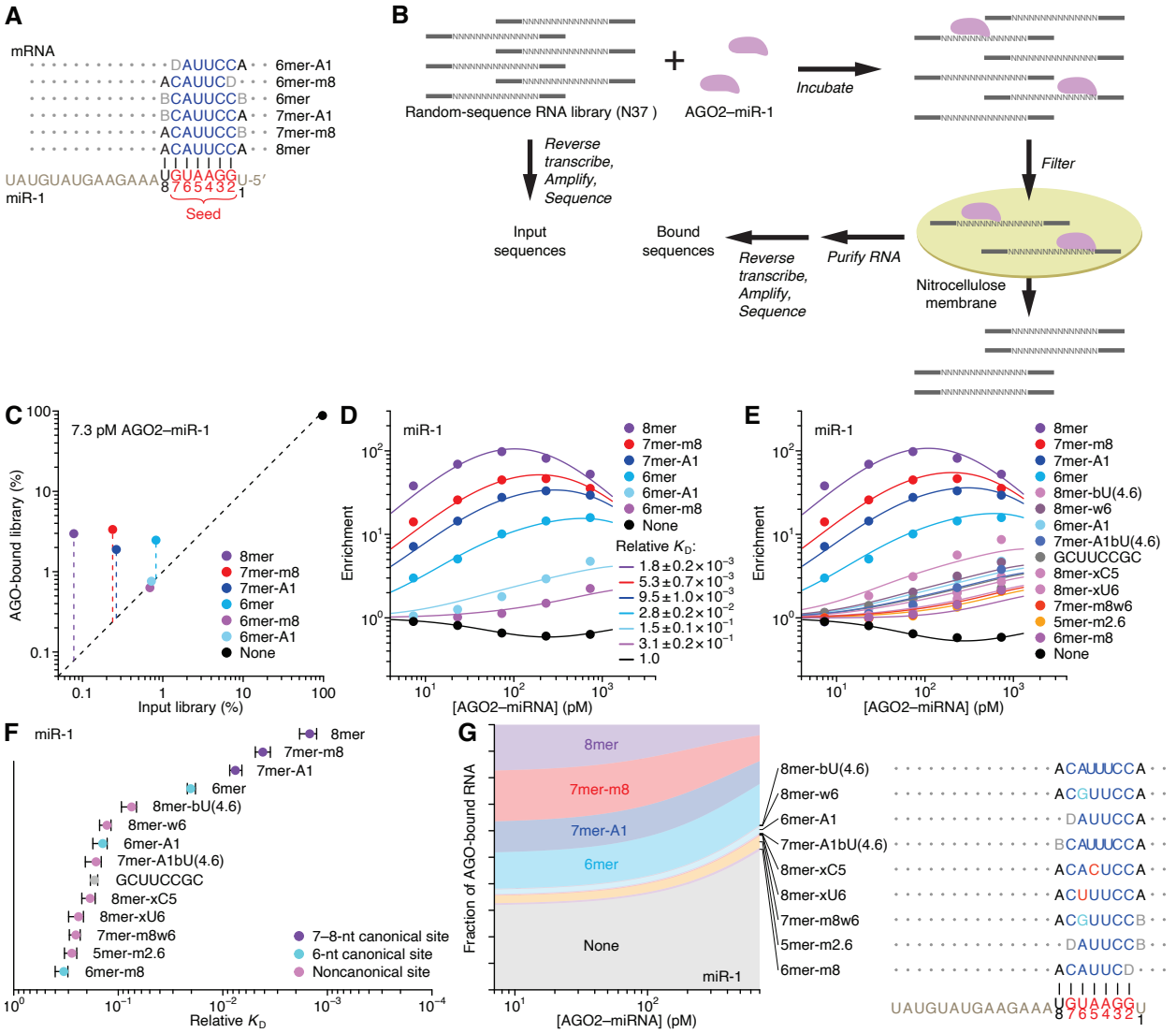


Figure 1. AGO-RBNS reveals binding affinities of canonical and previously uncharacterized miR-1 target sites.

(**A**) Canonical sites of miR-1. These sites have contiguous pairing (blue) to the miRNA seed (red), and some include an additional match to miRNA nucleotide 8 or an A opposite miRNA nucleotide 1 (B represents C, G, or U; D represents A, G, or U). (**B**) AGO-RBNS. Purified AGO2–miR-1 is incubated with excess RNA library molecules that each have a central block of 37 random-sequence positions (N37). After reaching binding equilibrium, the reaction is applied to a nitrocellulose membrane and washed under vacuum to separate library molecules bound to AGO2–miR-1 from those that are unbound. Molecules retained on the filter are purified, reverse transcribed, amplified, and sequenced. These sequences are compared with those generated directly from the input RNA library. (**C**) Enrichment of reads containing canonical miR-1 sites in the 7.3 pM AGO2–miR-1 library. Shown is the abundance of reads containing the indicated site (key) in the bound library plotted as a function of the respective abundance in the input library. Dashed vertical lines depict the enrichment in the bound library; dashed diagonal line shows $y = x$. Reads containing multiple sites were assigned to the site with greatest enrichment. (**D**) AGO-RBNS profile of the canonical miR-1 sites. Plotted is the enrichment of reads with the indicated canonical site (key) observed at each of the five AGO2–miR-1 concentrations of the AGO-

RBNS experiment, determined as in (C). Points show the observed values, and lines show the enrichment predicted from the mathematical model fit simultaneously to all of the data. Also shown for each site are K_D values obtained from fitting the model, listing the geometric mean ± the 95% confidence interval determined by resampling the read data, removing data for one AGO–miR-1 concentration and fitting the model to the remaining data, and repeating this procedure 200 times (40 times for each concentration omitted). (**E**) AGO-RBNS profile of the canonical and the newly identified noncanonical miR-1 sites (key). Sites are listed in the order of their K_D values and named and colored based on the most similar canonical site, indicating differences from this site with b (bulge), w (G–U wobble), or x (mismatch) followed by the nucleotide and its position. For example, the 8mer-bU(4.6) resembles a canonical 8mer site but has a bulged U at positions that would normally pair to miRNA nucleotides 4, 5, or 6. Everything else is the same as in (D). (**F**) Relative K_D values for the canonical and the newly identified noncanonical miR-1 sites determined in (E). Sites are classified as either 7–8-nt canonical sites (purple), 6-nt canonical sites (cyan), noncanonical sites (pink), or a sequence motif with no clear complementarity to miR-1 (gray). The solid vertical line marks the reference K_D value of 1.0 assigned to reads lacking an annotated site. Error bars indicate 95% confidence interval on the geometric mean, as in (D). (**G**) The proportion of AGO2–miR-1 bound to each site type. Shown are proportions inferred by the mathematical model over a range of AGO2‑miR-1 concentrations spanning the five experimental samples, plotted in the order of site affinity (top to bottom), using the same colors as in (E). On the right is the pairing of each noncanonical site, diagrammed as in (A), indicating Watson–Crick pairing (blue), wobble pairing (cyan), mismatched pairing (red), bulged nucleotides (compressed rendering), and terminal noncomplementarity (gray; B represents C, G, or U; D represents A, G, or U; H represents A, C, or U; V represents A, C, or G). The GCUUCCGC motif is omitted because it did not match miR-1 and did not mediate repression by miR-1 (Figure S5B).

nitrocellulose filter binding, reasoning that the rapid wash step of filter binding would improve retention of low-affinity molecules that would otherwise be lost during the wash steps of a pull-down. This modified method was highly reproducible, with high correspondence observed between the 9-nt *k*-mer enrichments of two independent experiments using different preparations of both AGO2–miR-1 and the RNA library (Figure S1B; $r^2 = 0.86$).

When analyzing our AGO-RBNS results, we first examined enrichment of the canonical miR-1 sites, comparing the frequency of these sites in RNA bound in the 7.3 pM AGO2–miR-1 sample with that of the input library. As expected from the site hierarchy observed in meta-analyses of site conservation and endogenous site efficacy (Bartel, 2009), the 8mer site (perfect match to miR-1 nucleotides 2–8 followed by an A) was most enriched (38-fold), followed by the

7mer-m8 site, then the 7mer-A1 site, and the 6mer site (Figures 1A and 1C). Little if any enrichment was observed for either the 6mer-A1 site or the 6mer-m8 site at this lowest concentration of 7.3 pM AGO2–miR-1 (Figures 1A and 1C), consistent with their weak signal in previous analyses of conservation and efficacy (Agarwal et al., 2015; Friedman et al., 2009; Kim et al., 2016). Enrichment of sites was quite uniform across the random-sequence region, which indicated minimal influence from either the primer-binding sequences or supplementary pairing to the 3′ region of the miRNA (Figure S1D). Although sites with supplementary pairing can have enhanced efficacy and affinity (Bartel, 2009; Brennecke et al., 2005; Wee et al., 2012), the minimal influence of supplementary pairing reflected the rarity of such sites in our library.

Analysis of enrichment of the six canonical sites across all five AGO2–miR-1 concentrations illustrated two hallmarks of this experimental platform (Lambert et al., 2014). First, as the concentration increased from 7.3–73 pM, enrichment for each of the six site types increased (Figure 1D), which was attributable to an increase in signal over a constant low background of library molecules isolated even in the absence of AGO2–miR-1. Second, as the AGO2–miR-1 concentration increased beyond 73 pM, 8mer enrichment decreased, and at the highest AGO2–miR-1 concentration, enrichment of the 7mer-m8 and 7mer-A1 site decreased (Figure 1D). These waning enrichments indicated the onset of saturation for these high-affinity sites (Lambert et al., 2014). These two features, driven by AGO–miRNA-independent background and partial saturation of the higher-affinity sites, respectively, caused differences in enrichment values for different site types to be highly dependent on the AGO2–miR-1 concentration; the lower AGO2–miR-1 concentrations provided greater discrimination between the higher-affinity site types, the higher AGO2–miR-1 concentrations provided greater discrimination between the lower-affinity site types, and no single concentration provided results that quantitatively reflected differences in relative binding affinities.

To account for background binding and ligand saturation, we developed a computational strategy that simultaneously incorporated information from all concentrations of an RBNS experiment to calculate relative K_D values. Underlying this strategy was an equilibrium-binding model that predicts the observed enrichment of each site type across the concentration series as a function of the K_D values for each miRNA site type (including the "no-site" type), as well as the stock concentration of purified AGO2–miR-1 and a constant amount of library recovered as background in all samples. Using this model, we performed maximum likelihood estimation (MLE) to fit the relative K_D values, which explained the observed data well (Figure 1D). Moreover, these relative K_D values were robustly estimated, as indicated by comparing values obtained using results from only four of the five AGO2–miR-1 concentrations ($r^2 \geq 0.994$ for each of the ten pairwise comparisons; Figures S1F and S1G). These quantitative binding affinities followed the same hierarchy as observed for site enrichment, but the differences in affinities were of greater magnitude (Figures 1D and S1C).

Up to this point, our analysis was informed by the wealth of previous computational and experimental data showing the importance of a perfect 6–8-nt match to the seed region (Bartel, 2009). However, the ability to calculate the relative K_D of any *k*-mer of length ≤12 nt (the 12-nt limit imposed by the sparsity of reads with longer *k*-mers) provided the opportunity for a de novo search for sites, without bias from any previous knowledge. In this search, we 1) calculated the enrichment of all 10-nt *k*-mers in the bound RNA in the 730 pM AGO2–miR-1 sample, which was the sample with the most sensitivity for detecting low-affinity sites, 2) determined the extent of complementarity between the ten most enriched *k*-mers and the miR-1 sequence, 3) assigned a site most consistent with the observed *k*-mers, and 4) removed all reads containing this newly identified site from both the bound and input libraries. These four steps were iterated until no 10-nt *k*-mer remained that was enriched ≥10-fold, thereby generating 14 sites for AGO2–miR-1. We

then applied our MLE procedure to calculate relative K_D values for this expanded list of sites (Figures 1E and 1F).

This unbiased approach demonstrated that the 8mer, 7mer-m8, 7mer-A1, and 6mer sites to miR-1 were the highest-affinity site types of lengths ≤10 nt. It also identified eight novel sites with binding affinities resembling those of the 6mer-m8 and the 6mer-A1 (Figure 1F). Comparison of these sites to the sequence of miR-1 revealed that miR-1 can tolerate either a wobble G at position 6 or a bulged U somewhere between positions 4 and 6 and achieve affinity at least 7–11-fold above that of the remaining no-site reads, and that it can tolerate either a mismatched C at position 5 or a mismatched U at position 6 and achieve affinity 4–5-fold above that of the no-site reads. The GCUUCCGC motif also passed our cutoffs, which was more difficult to explain, because it had contiguous complementarity to positions 2–5 of miR-1 flanked by noncomplementary GC dinucleotides on both sides. Nonetheless, among the 1,398,100 possible motifs ≤10 nt, this was the only one that satisfied our criteria yet was difficult to attribute to miRNA pairing.

Our analytical approach and its underlying biochemical model also allowed us to infer the proportion of AGO2–miR-1 bound to each site (Figure 1G). The 8mer site occupied 3.8–17% of the silencing complex over the concentration course, whereas the 7mer-m8, by virtue of its greater abundance, occupied a somewhat greater fraction of the complex. In aggregate, the marginal sites—including the 6mer-A1, 6mer-m8, and seven noncanonical sites—occupied 6.1–9.8% of the AGO2–miR-1 complex. Moreover, because of their very high abundance, library molecules with no identified site occupied 32–53% of the complex (Figure 1G). These results support the inference that the summed contributions of background binding and low-affinity sites to intracellular AGO occupancy is of the same order of magnitude as that of canonical sites, suggesting that an individual AGO–miRNA complex spends about half its time associated with a

vast repertoire of background and low-affinity sites (Denzler et al., 2014, 2016). This phenomenon would help explain why sequences without recognizable sites often crosslink to AGO in cells.

Our results confirmed that AGO2–miR-1 binds the 8mer, 7mer-m8, 7mer-A1, and 6mer sites most effectively and revealed the relative binding affinities and occupancies of these sites. In addition, our results uncovered weak yet specific affinity to the 6mer-A1 and 6mer-m8 sites plus seven noncanonical sites, all with affinities outside the dynamic range of recent high-throughput imaging experiments (Becker et al., 2019). Although alternative binding sites for miRNAs have been proposed based on high-throughput in vivo crosslinking studies (Chi et al., 2012; Grosswendt et al., 2014; Helwak et al., 2013; Khorshid et al., 2013; Loeb et al., 2012), our approach provided quantification of the relative strength of these sites without the confounding effects of differential crosslinking efficiencies, potentially enabling their incorporation into a quantitative framework of miRNA targeting.

Distinct canonical and noncanonical binding of different miRNAs

We extended our analysis to five additional miRNAs, including let-7a, miR-7, miR-124, and miR-155 of mammals, chosen for their sequence conservation as well as the availability of data examining their regulatory activities, intracellular binding sites, or in vitro binding affinities (Bartel, 2018; Chi et al., 2012; Loeb et al., 2012; Salomon et al., 2015; Wee et al., 2012), and lsy-6 of nematodes, which is thought to bind unusually weakly to its canonical sites (Garcia et al., 2011) (Figures 2, S2B, and S2C). In the case of let-7a, previous biochemical analyses have determined the K_D values of some canonical sites (Becker et al., 2019; Salomon et al., 2015; Wee et al., 2012), and our values agreed well, which further validated our high-throughput approach (Figure S1H).

The site-affinity profile of let-7a resembled that of miR-1, except the 6mer-m8 and 6mer-A1 sites for let-7a had greater binding affinity than essentially all of the noncanonical sites (Figure 2A). As with miR-1, the noncanonical sites each paired to the seed region but did so imperfectly, typically with a single wobble, single mismatch, or single-nucleotide bulge, but these imperfections differed from those observed for miR-1 (Figures 1F and 2A).

The site-affinity profiles of miR-124, miR-155, lsy-6, and miR-7 resembled those of miR-1 and let-7a. All but one included the six canonical sites (with miR-7 missing the 6mer-m8 site), and all contained noncanonical sites with extensive yet imperfect pairing to the miRNA seeds, the imperfections tending to occur at different positions and with different mismatched- or bulged-nucleotide identities for different miRNAs, (Figures 2B, 2C, S2B, and S2C). In contrast to the noncanonical sites of miR-1 and let-7a, more of the noncanonical sites of the other four miRNAs had affinities interspersed with those of the top four canonical sites. Moreover, the profiles for miR-155, miR-124, and lsy-6 also included sites with extended (9–11-nt) complementarity to the miRNA 3′ region. These sites had estimated K_D values that were derived from reads with little more than chance complementarity to the miRNA seed, and they had uniform enrichment across the length of the random-sequence region (Figure S1E), which indicated that these sites represented an alternative binding mode dominated by extensive pairing to the 3′ region without involvement of the seed region (Figures 2B, 2C, and S2B). We named them “3′-only sites.”

In some respects, the 3′-only sites resembled noncanonical sites known as centered sites, which are reported to function in mammalian cells (Shin et al., 2010). Like 3′-only sites, centered sites have extensive perfect pairing to the miRNA, but for centered sites, this pairing begins at miRNA positions 3 or 4 and extends 11–12 nt through the center of the miRNA (Shin et al., 2010). Our unbiased search for sites did not identify centered sites for any of the six miRNAs.

We therefore directly queried the region of each miRNA to which extensive noncanonical pairing was favored, determining the affinity of sequences with 11-nt segments of perfect complementarity to the miRNA sequence, scanning from miRNA position 3 to the 3′ end of the miRNA (Figure 3A). For miR-155, miR-124, and lsy-6, sequences with 11-nt sites that paired to the miRNA 3′ region bound with greater affinity than did those with a canonical 6mer site, whereas for let-7a and miR-1, and miR-7, none of the 11-nt sites conferred stronger binding than did the 6mer. Moreover, for all six miRNAs, the 11-nt sites that satisfied the criteria for annotation as centered sites conferred binding ≤2-fold stronger than that of the 6mer-m8 site, which also starts at position 3 but extends only 6 nt. These results called into question the function of centered sites, although we cannot rule out the possibility that centered sites are recognized by some miRNAs and not others. Indeed, the newly identified 3′-only sites functioned for only miR-155, miR-124, and lsy-6, and even among these, the optimal region of pairing differed, occurring at positions 13–23, 9–19, and 8–18, respectively (Figure 3A).

When evaluating other types of noncanonical sites proposed to confer widespread repression in mammalian cells (Chi et al., 2012; Kim et al., 2016), we found that all but two bound with affinities difficult to distinguish from background. One of these two was the 5-nt site matching miRNA positions 2–6 (5mer-m2.6) (Kim et al., 2016), which was bound by miR-1, let-7a, and miR-7 but not by the other three miRNAs (Figure S3). The other was the pivot site (Chi et al., 2012), which was bound by miR-124 [e.g., 8mer-bG(6.7); Figure 2C] and lsy-6 [e.g., 8mer-bA(6.7); Figure S2B] but not by the other four miRNAs (Figure S4). The absence of a pivot site for let-7a in our data contrasted with the prior results, in which the pivot site was reported for both miR-124, let-7a, and miR-708 (Chi et al., 2012). However, our results are consistent with those of another high-throughput study, which reports weak affinity for this site when measured within 32 different target sequence contexts (Becker et al., 2019). More

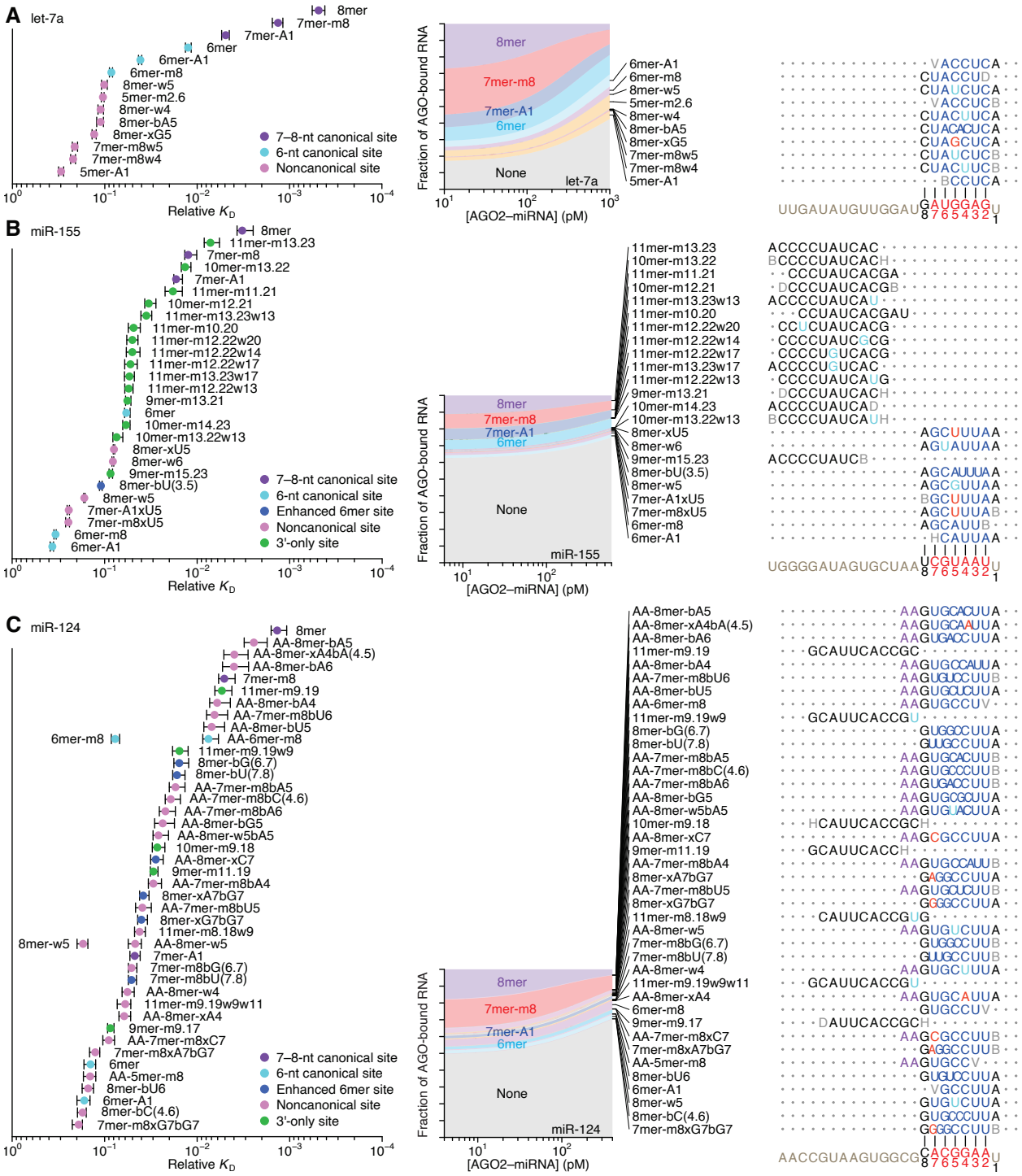


Figure 2. Distinct canonical and noncanonical binding of different miRNAs.
(**A**–**C**) Relative K_D values and proportional occupancy of established and newly identified sites of let-7a (A), miR-155 (B), and miR-124 (C). The two miR-124 sites that were present as a 5′-AA–extended form in addition to an unextended form are shown on the same line (C). Relative K_D values are plotted as in Figure 1F but in some cases with additional categories, either for 3′-only sites (green) (B and C) or for 6-nt canonical sites enhanced by either additional wobble-pairing or additional Watson–Crick complementarity separated by a bulged nucleotide (blue) (B

and C). The proportion of AGO2–miRNA bound to each site type is estimated and shown as in Figure 1G. These analyses also detected a GCACUUUA motif for let-7a and AACGAGGA motif for miR-155, which were assigned relative K_D values of $7.1 \pm 0.8 \times 10^{-2}$ and $6 \pm 1 \times 10^{-2}$, respectively. These motifs are excluded because each did not match its respective miRNA and did not mediate repression by its respective miRNA (Figure S5B).

generally, these two previously identified noncanonical site types resembled the newly identified noncanonical sites with extensive yet imperfect pairing to the seed region, in that they function for only a limited number of miRNAs.

In addition to the differences in noncanonical site types observed for each miRNA, we also observed pronounced miRNA-specific differences in the relative affinities of the canonical site types. For example, for miR-155, the affinity of the 7mer-A1 nearly matched that of the 7mer-m8, whereas for miR-124, the affinity of the 7mer-A1 was >9-fold lower than that of the 7mer-m8. These results implied that the relative contributions of the A at target position 1 and the match at target position 8 can substantially differ for different miRNAs. Prior studies show that AGO proteins remodel the thermodynamic properties of their loaded RNA guides (Salomon et al., 2015; Wee et al., 2012), and our results show that the sequence of the guide strongly influences the nature of this remodeling, leading to differences in relative affinities across canonical site types and a distinct repertoire of noncanonical site types for each miRNA.

The energetics of canonical binding

With the relative K_D values for the canonical binding sites of six miRNAs in hand, we examined the energetic relationship between the A at target position 1 (A1) and the match at miRNA position 8 (m8), within the framework analogous to a double-mutant cycle (Figure 3B, left). The apparent binding-energy contributions of the m8 and A1 ($\Delta\Delta G_{m8}$ and $\Delta\Delta G_{A1}$, respectively) were largely independent, as inferred from the relative K_D values of the four site types. That is, for

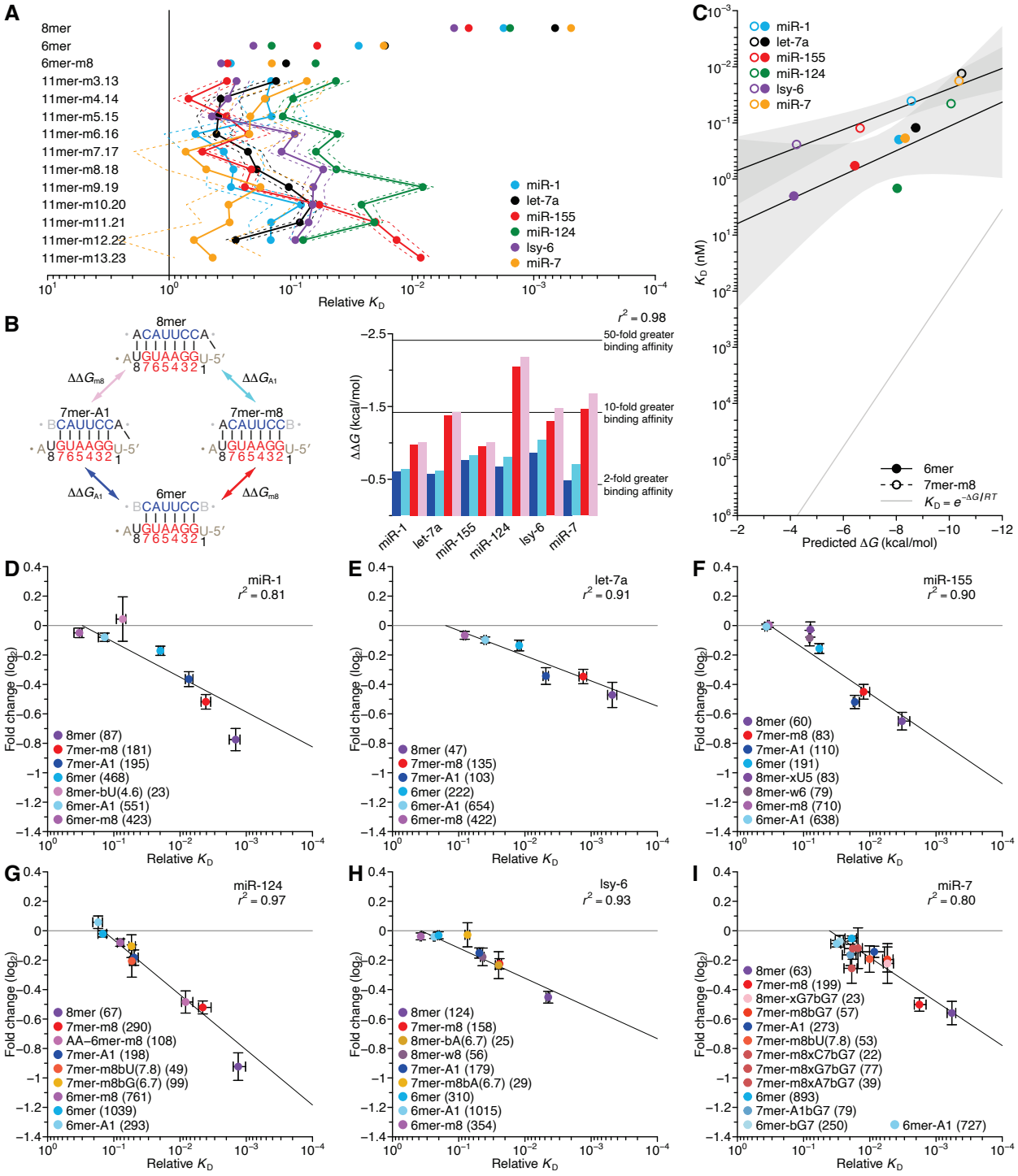


Figure 3. Additional analyses of binding affinities and the correspondence between binding affinity and repression efficacy.

(**A**) Diverse functionality and position dependence of 11-nt 3′-only sites. Relative K_D values for each potential 11-nt 3′-only site are plotted for the indicated miRNAs (key). For reference, values for the 8mer, 6mer, and 6mer-m8 sites are also plotted. The solid vertical line marks the reference K_D value of 1.0, as in Figure 1F. The solid and dashed lines indicate geometric mean and 95% confidence interval, respectively, determined as in Figure 1D. (**B**) The independent

contributions of the A1 and m8 features. On the left, a double-mutant cycle depicts the affinity differences observed among the four top canonical sites for miR-1, as imparted by the independent contributions of the A1 and m8 features and their potential interaction. On the right, the apparent binding contributions of the A1 ($\Delta\Delta G_{A1}$, blue and cyan) or m8 ($\Delta\Delta G_{m8}$, red and pink) features are plotted, determined from the ratio of relative K_D values of either the 7mer-A1 and the 6mer (blue), the 8mer and the 7mer-m8 (cyan), the 7mer-m8 and the 6mer (red), or the 8mer and the 7mer-A1 (pink) for the indicated AGO2–miRNA complexes. The r^2 reports on the degree of $\Delta\Delta G$ similarity for both the m8 and A1 features using either of the relevant site-type pairs across all six complexes. (**C**) The relationship between the observed relative K_D values and predicted pairing stability of the 6mer (filled circles) and 7mer-m8 (open circles) sites of the indicated AGO–miRNA complex (key), under the assumption that the K_D value for library molecules without a site was 10 nM for all AGO–miRNA complexes. The two black lines are the best fit of the relationship observed for each of the site types (gray regions, 95% confidence interval). The gray line shows the expected relationship with the predicted stabilities given by $K_D = e^{-\Delta G/RT}$. (**D**–**I**) The relationship between repression efficacy and relative K_D values for the indicated sites of miR-1 (D), let-7a (E), miR-155 (F), miR-124 (G), lsy-6 (H), and miR-7 (I). The number of sites of each type in the 3′ UTRs is indicated (parentheses). To include information from mRNAs with multiple sites, multiple linear regression was applied to determine the log fold-change attributable to each site type (error bars, 95% confidence interval). The relative K_D values are those of Figures 1, 2, and S2 (error bars, 95% confidence interval). Lines show the best fit to the data, determined by least-squares regression, weighting residuals using the 95% confidence intervals of the log fold-change estimates. The r^2 values were calculated using similarly weighted Pearson correlations.

each miRNA, the $\Delta\Delta G_{m8}$ inferred in presence of the A1 (using the ratio of the 8mer and 7mer-A1 K_D values) resembled that inferred in the absence of the A1 (using the ratio of the 7mer-m8 and 6mer K_D values), and vice versa (Figure 3B).

The relative K_D values for canonical sites of six miRNAs provided the opportunity to examine the relationship between the predicted free energy of site pairing and measured site affinities. We focused on the 6mer and 7mer-m8 sites, because they lack the A1, which does not pair to the miRNA (Figure 1A) (Lewis et al., 2005; Schirle et al., 2015). Consistent with the importance of base pairing for site recognition and the known relationship between predicted seed-pairing stability and repression efficacy (Garcia et al., 2011), affinity increased with increased predicted pairing stability, although this increase was statistically significant for only the 7mer-m8 site type (Figure 3C; $p = 0.09$ and 0.005 for the 6mer and 7mer-m8 sites,

respectively). However, for both site types, the slope of the relationship was significantly less than expected from $K_D = e^{-\Delta G/RT}$, where ΔG is the change in free energy, R is the universal gas constant, and T is temperature ($p = 0.008$ and 8×10^{-5}, respectively). When considered together with previous analysis of a miRNA with enhanced seed pairing stability, these results indicated that in remodeling the thermodynamic properties of the loaded miRNAs, AGO not only enhances the affinity of seed-matched interactions but also dampens the intrinsic differences in seed-pairing stabilities that would otherwise impose much greater inequities between the targeting efficacies of different miRNAs (Salomon et al., 2015). Thus, although lsy-6, which has unusually poor predicted seed-pairing stability (Garcia et al., 2011), did indeed have the weakest site-binding affinity of the six miRNAs, the difference between its binding affinity and that of the other miRNAs was less than might have been expected.

Correspondence with repression observed in the cell

To evaluate the relevance of our in vitro binding results to intracellular miRNA-mediated repression, we examined the relationship between the relative K_D measurements and the repression of endogenous mRNAs after miRNA transfection into HeLa cells. When examining intracellular repression attributable to 3′-UTR (3′ untranslated region) sites of the transfected miRNA, we observed a pronounced relationship between AGO-RBNS–determined K_D values and mRNA fold changes (Figures 3D–3I; $r^2 = 0.80$–0.97). For instance, the different relative affinities of the 7mer-A1 and 7mer-m8 sites, most extremely observed for sites of miR-155 and miR-124, was nearly perfectly mirrored by the relative efficacy of these sites in mediating repression in the cell (Figures 3F and 3G). A similar correspondence between relative K_D values and repression was observed for the noncanonical sites that had both sufficient affinity and sufficient representation in the HeLa transcriptome to be evaluated using this analysis (Figures

3D–3I). These included the pivot sites for miR-124 and lsy-6 and the bulge-G7–containing sites for miR-7 (Figures 3G and 3I).

Analysis of mRNA changes following miRNA transfection was not suitable for measuring efficacy of the highest-affinity noncanonical sites because these sites lacked sufficient representation in endogenous 3′ UTRs. Therefore, we implemented a massively parallel reporter assay designed to examine the efficacy of every site type identified by AGO-RBNS, each in 184 different 3′ UTR sequence contexts (Figure S5A). This assay showed that 3′-only sites and other high-affinity-but-rare noncanonical site types do mediate repression in cells and that their efficacies tend to track with their affinities (Figure S5B). In sum, we found a strong correspondence between intracellular repression and in vitro binding affinity, regardless of miRNA identity and regardless of whether the target site is canonical or noncanonical or within an endogenous or a reporter mRNA. This result supported a model in which repression is a function of miRNA occupancy, as dictated by site affinity, and thus miRNA- and site-specific differences in binding affinities explain substantial differences in repression.

The strong influence of flanking dinucleotide sequences

AU-rich nucleotide composition immediately flanking miRNA sites has long been associated with increased site conservation and efficacy in cells (Grimson et al., 2007; Lewis et al., 2005; Nielsen et al., 2007), but the mechanistic basis of this phenomenon had not been investigated, presumably because of the sparsity of affinity measurements. The AGO-RBNS data provided the means to overcome this limitation. We first separated the miR-1 8mer site into 256 different 12-nt sites, on the basis of the dinucleotide sequences immediately flanking each side of the 8mer, and determined relative K_D values for each (Figure 4A). This analysis revealed a ~100-fold range in values, depending on the identities of the flanking dinucleotides, with binding affinity strongly

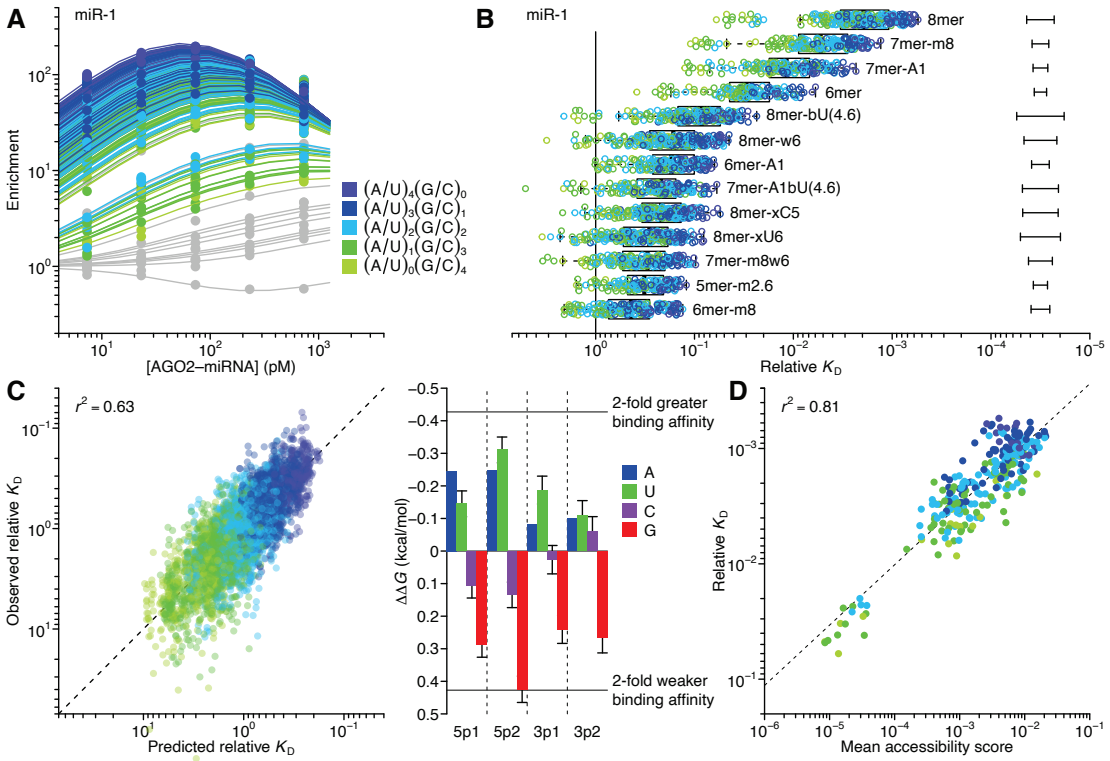


Figure 4. The influence of flanking dinucleotide sequence context.
(**A**) AGO-RBNS profile of miR-1 sites, showing results for the 8mer separated into 256 different 12-nt sites on the basis of the identities of the two dinucleotides immediately flanking the 8mer. For each 12-nt site, the points and line are colored on the basis of the AU content of the flanking dinucleotides (key). For context, results of Figure 1E are replotted in gray. Everything else is the same as in Figure 1E. (**B**) Relative K_D values for each miR-1 site identified in Figure 1F separated into 144 to 256 sites as in (A) on the basis of the identities of the flanking dinucleotides. The points are colored as in (A). Error bars indicate median 95% confidence interval across all K_D values. Everything else is the same as in Figure 1F. (**C**) Consistency of flanking-dinucleotide effect across miRNA and site type. At the left is a comparison of observed relative K_D values and results of a mathematical model that used multiple linear regression to predict the influence of flanking dinucleotides. Plotted are results for all flanking dinucleotide contexts of all six canonical site types, for all six miRNAs, normalized to the average affinity of each canonical site. Predictions of the model are those observed in a sixfold cross-validation, training on the results for five miRNAs and reporting the predictions for the held-out miRNA. The points for five outliers are not shown. The r^2 quantifies the agreement between the predicted and actual values, considering all points. On the right, the model coefficients (multiplied by $-RT$, where $T = 310.15$ K) corresponding to each of the four nucleotides of the 5′ (5p) and 3′ (3p) dinucleotides in the 5′-to-3′ direction are plotted (error bars, 95% confidence interval). (**D**) Relationship between the mean structural-accessibility score and the relative K_D for the 256 12-nt sites containing the miR-1 8mer flanked by each of the dinucleotide combinations. Points are colored as in (A). Linear regression (dashed line) and calculation of r^2 were performed using log-transformed values. For an analysis of the relationship between 8mer flanking-dinucleotide K_D and structural accessibility over a range of window lengths and positions relative to the 8mer site, see Figure S6G.

tracking the AU content of the flanking dinucleotides. Extending this analysis across all miR-1 site types (Figure 4B), as well as to sites to the other five miRNAs (Figures S6A–S6E), yielded similar results. The effect of flanking-dinucleotide context was of such magnitude that it often exceeded the affinity differences observed between miRNA-site types. Indeed, for each miRNA, at least one 6-nt canonical site in its most favorable context had greater affinity than that of the 8mer site in its least favorable context (Figures 4B, and S6A–S6E).

To identify general features of the flanking-dinucleotide effect across miRNA sequences and site types, we trained a multiple linear-regression model on the complete set of flanking-dinucleotide K_D values corresponding to all six canonical site types of each miRNA, fitting the effects at each of the four positions within the two flanking dinucleotides. The output of the model agreed well with the observed K_D values (Figure 4C, left; $r^2 = 0.63$), which indicated that the effects of the flanking dinucleotides were largely consistent between miRNAs and between site types of each miRNA. The output of the model also corresponded with the efficacy of intracellular repression, which indicated that these effects on K_D values were consequential in cells (Figure S6F). A and U nucleotides each enhanced affinity, whereas G nucleotides reduced affinity, and C nucleotides were intermediate or neutral (Figure 4C, right). Moreover, the identity of the 5′ flanking dinucleotide, which must come into close proximity with the central RNA-binding channel of AGO (Schirle et al., 2014), contributed more to binding affinity than did the 3′ flanking sequence (Figure 4C, right).

One explanation for this hierarchy of flanking nucleotide contributions, with A ≈ U > C > G, is that it inversely reflected the propensity of these nucleotides to stabilize RNA secondary structure that could occlude binding of the silencing complex (Ameres et al., 2007; Brown et al., 2005; Chen et al., 2009; Kedde et al., 2007, 2010; Kertesz et al., 2007; Obernosterer, 2006; Rudnick et al., 2008; Tafer et al., 2008). To investigate this potential role for structural

accessibility in influencing binding, we compared the predicted structural accessibility of 8mer sites in the input and bound libraries of the AGO2–miR-1 experiment, using a score for predicted structural accessibility previously optimized on data examining miRNA-mediated repression (Agarwal et al., 2015; Tafer et al., 2008). This score is based on the predicted probability that the 14-nt segment at target positions 1–14 is unpaired. We found that predicted accessibilities of sites in the bound libraries were substantially greater than those for sites in the input library and that the difference was greatest for the samples with the lower AGO2–miR-1 concentrations (Figure S6G), as expected if the accessibility score was predictive of site accessibility and if the most accessible sites were the most preferentially bound.

To build on these results, we examined the relationship between predicted structural accessibility and binding affinity for each of the 256 flanking dinucleotide possibilities. For each input read with a miR-1 8mer site, the accessibility score of that site was calculated. The sites were then differentiated on the basis of their flanking dinucleotides into 256 12-nt sites, and the geometric mean of the structural-accessibility scores of each of these extended sites was compared with the AGO-RBNS–derived relative K_D value (Figures 4D and S6H). A notable correlation was observed ($r^2 = 0.82$, $p < 10^{-15}$), with all 16 sites containing a 5′-flanking GG dinucleotide having both unusually poor affinities and unusually low accessibility scores. Moreover, sampling reads from the input library to match the predicted accessibility of sites in the bound library recapitulated the flanking dinucleotide preferences observed in the bound library (Figure S6I; $r^2 = 0.79$). Taken together, our results demonstrate that local sequence context has a large influence on miRNA–target binding affinity and indicate that this influence results predominantly from the differential propensities of flanking sequences to favor structures that occlude site accessibility.

A biochemical model predictive of miRNA-mediated repression

Inspired by the finding that measured affinities strongly corresponded to the repression observed in cells (Figures 3D–3I), we set out to build a biochemical framework that predicts the degree to which a miRNA represses each mRNA. Biochemical principles have been used to model miR-21–directed mRNA slicing (Becker et al., 2019). However, previous efforts that used biochemical principles to model aspects of the predominant mode of miRNA-mediated repression, including competition between endogenous target sites (Bosson et al., 2014; Denzler et al., 2016; Jens and Rajewsky, 2014) and the influence of miRNAs on reporter gene–expression noise (Schmiedel et al., 2015), were severely limited by the sparsity of the data. Our ability to measure the relative binding affinity of a miRNA to any 12-nt sequence enabled modeling of the quantitative effects of the six miRNAs on each cellular mRNA.

We first re-analyzed all six AGO-RBNS experiments to calculate, for each miRNA, the relative K_D values for all 262,144 12-nt *k*-mers that contained at least four contiguous nucleotides of the canonical 8mer site (Figure 5A). These potential binding sites included the canonical sites and most of the noncanonical sites that we had identified, each within a diversity of flanking sequence contexts (Figures 1F and 2). For each mRNA *m* and transfected miRNA *g*, the steady-state occupancy $N_{m,g}$ (i.e., the average number of AGO–miRNA complexes loaded with miRNA *g* bound to mRNA *m*) was predicted as a function of the K_D values of the potential binding sites contained within the mRNA open reading frame (ORF) and 3′ UTR, as well as the concentration of the unbound AGO–miRNA$_g$ complex a_g, which was fit as a single value for each transfected miRNA (Figure 5B, equation 1). This occupancy value enabled prediction of a biochemically informed expectation of repression, assuming that the added effect of the miRNA on the basal decay rate scaled with the basal rate and $N_{m,g}$ (Figure 5B, equation 2). To isolate the effects of a transfected miRNA over background, we further offset our prediction of repression by a

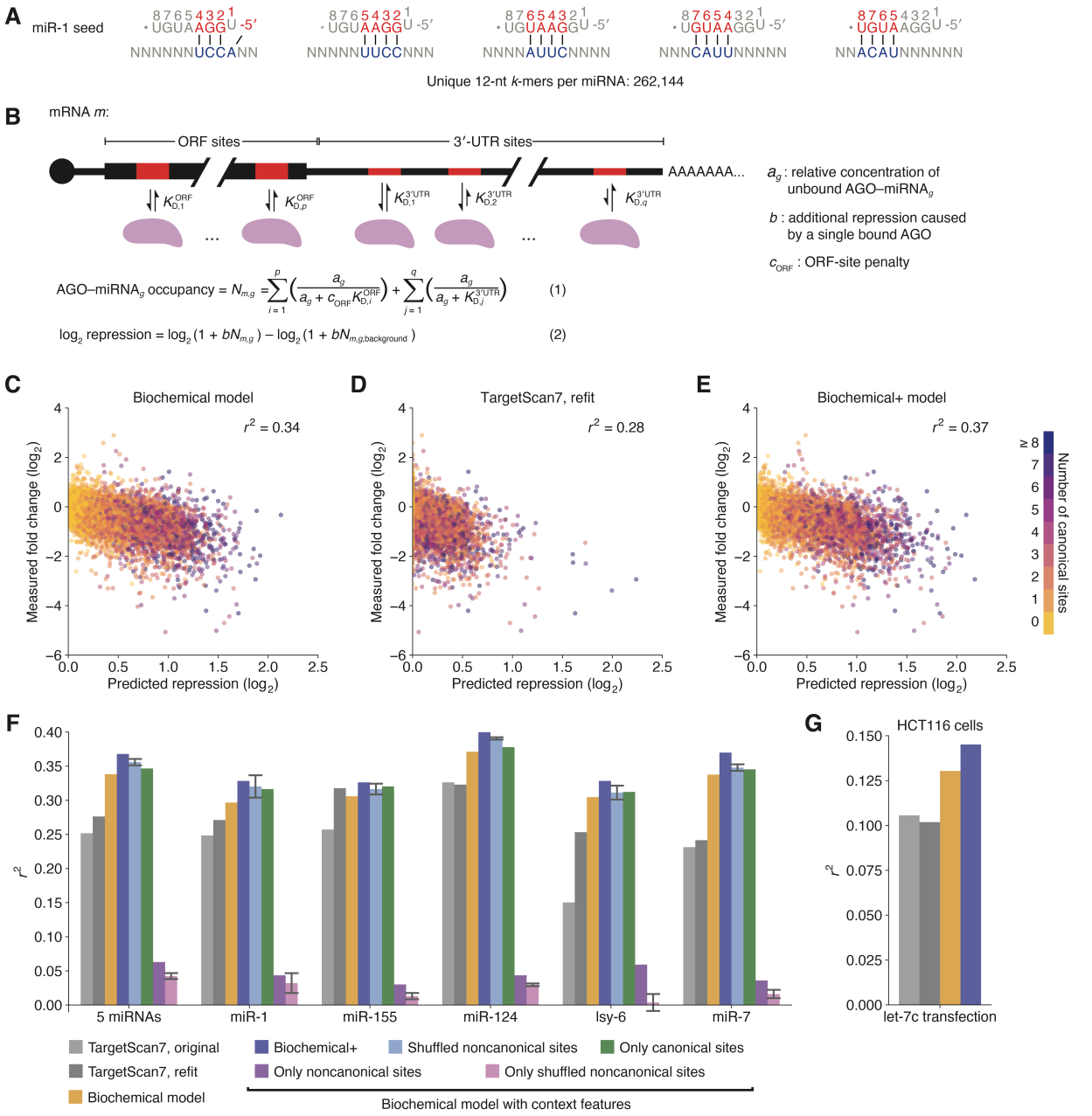


Figure 5. AGO-RBNS K_D values enable a predictive model of miRNA-mediated repression in cells.

(**A**) The 262,144 12-nt *k*-mers with at least four contiguous matches to the extended seed region of miR-1, for which relative K_D values were determined. Relative K_D values were similarly determined for the analogous *k*-mers of the other five miRNAs. (**B**) Biochemical model for estimating miRNA-mediated repression of an mRNA using the relative K_D values of the 12-nt *k*-mers in the mRNA. (**C**) Performance of the biochemical model as evaluated using the combined results of five miRNAs. Plotted is the relationship between mRNA changes observed after transfecting a miRNA and those predicted by the model. Each point represents the mRNA from one gene after transfection of a miRNA and is colored according to the number of canonical sites in the mRNA 3′ UTR (key). For easier visual comparison between mRNAs, *y*-axis points for the

same mRNA are adjusted by the extrapolated expression level of the mRNA with no transfected miRNA. The Pearson's r^2 between measured and predicted values is for unadjusted values and is reported in the upper right. (**D**). Performance of the retrained TargetScan7 model. Everything else is the same as in (C). (**E**) Performance of the biochemical+ model. Everything else is the same as in (C). (**F**) Model performances and the contribution of cognate noncanonical sites to performance of the biochemical+ model. Results for each model (key) are plotted for individual miRNAs and for all five miRNAs combined (error bars, standard deviation). (**G**) Performances of models tested on mRNA changes observed after transfecting let-7c into HCT116 cells engineered to have reduced endogenous miRNA expression (Linsley et al., 2007). This analysis used the average a_g fit for the five miRNAs in (F). Everything else is the same as in (F).

background-binding term (Figure 5B, $N_{m,g,\text{background}}$). The calculation of predicted repression required an estimate of how much a single bound RISC complex affected the mRNA decay rate (Figure 5B, b), which was fit as a global value. Additionally, to account for the observation that sites in ORFs are less effective than those in 3′ UTRs (Bartel, 2009), our model included a penalty term for sites in ORFs, which was also fit as a global value (Figure 5B). Because no appreciable repression was observed from sites in 5′ UTRs, our model did not consider these sites.

Our biochemical model was fit against repression observed in HeLa cells transfected with one of five miRNAs with RBNS-derived measurements (let-7a was excluded because of its high endogenous expression in HeLa cells). A strong correspondence was observed when comparing mRNA changes measured upon miRNA transfection with those predicted by the model (Figure S7A; r^2 = 0.30–0.37).

The overall performance of our biochemical model (Figure 5C; r^2 = 0.34) exceeded those of the 30 target-prediction algorithms ($r^2 \leq 0.14$) that were also tested on changes in mRNA levels observed in response to miRNA transfection (Agarwal et al., 2015). We reasoned that in addition to our biochemical framework and the use of experimentally measured affinity values, other aspects of our analysis might have contributed to this improvement. For example, the

miRNAs chosen for RBNS have high efficacy in transfection experiments, and our RNA-sequencing (RNA-seq) datasets generally had stronger signal over background compared to microarray datasets used to train and test previous target-prediction algorithms. Indeed, when evaluated on the same five datasets, the performance of the latest TargetScan model (TargetScan7) improved from an r^2 of 0.14 to an r^2 of 0.25 (Figure S7B). To explore the possibility that TargetScan7 might also benefit from training on this type of improved data, we generated transfection datasets for 11 additional miRNAs and retrained TargetScan7 on the collection of 16 miRNA-transfection datasets (again omitting the let-7a dataset), putting aside one dataset each time in a 16-fold cross-validation. Training and testing TargetScan on improved datasets further increased the r^2 to 0.28 for the five miRNAs with AGO-RBNS data (Figure 5D). Nonetheless, the biochemical model still outperformed the retrained TargetScan by >20%, which showed that the use of measured affinity values in a biochemical framework substantially increased prediction performance.

Many features known to correlate with targeting efficacy were captured by our biochemical model. Indeed, the contribution of certain features, such as site type (Bartel, 2009), predicted seed-pairing stability (Garcia et al., 2011), and nucleotide identities at specific miRNA or site positions (Agarwal et al., 2015), are expected to be represented more accurately in the miRNA-specific K_D values of the 12-nt *k*-mers than when generalized across miRNAs. However, these K_D values did not fully capture other factors that that influence the affinity between miRNAs and their target sites in cells, including the structural accessibility of sites within their larger mRNA contexts and the contribution of supplementary pairing to the miRNA 3′ region, which influences approximately 5% of sites (Bartel, 2009). Without sufficient biochemical data quantifying these effects, we approximated their influence using scoring metrics known to correlate with miRNA targeting efficacy (Agarwal et al., 2015; Grimson et al., 2007) and

allowed them to modify the K_D values additively in log space (i.e., linearly in free-energy space). Incorporating each of these metrics slightly improved the performance of the biochemical model, as did incorporating a score for the evolutionary conservation of the site (Friedman et al., 2009), which helped account for additional unknown or imperfectly captured factors that influence targeting efficacy (Figure S7C). Simultaneously incorporating all three metrics to generate what we call the "biochemical+ model" improved the r^2 by 9% to 0.37 (Figure 5E).

To examine how well our models generalized to another cell type and to a miRNA family not used for fitting (let-7), we evaluated them on repression data collected after transfecting let-7c into HCT116 cells that had been engineered to not express endogenous miRNAs (Linsley et al., 2007). Although these data had a considerably lower signal-to-noise ratio, which lowered all r^2 values, our biochemical models substantially outperformed TargetScan7 (Figure 5G). This improvement extended to predicting repression after transfecting miR-124 and miR-7 into human embryonic kidney (HEK) 293 cells (Hausser et al., 2009) (Figure S8A). Additional analyses showed that the biochemical+ model performed at least as well as in vivo crosslinking (CLIP-seq) approaches in identifying the mRNAs most repressed upon miRNA transfection or most derepressed upon miRNA knockout (Hafner et al., 2010; Hausser et al., 2009; Loeb et al., 2012) (Figures S8B–S8D). Furthermore, for individual CLIP clusters enriched in wild type relative to miR-155 knockout, we observed a correlation between the occupancy predicted by our K_D values and the observed enrichment of the cluster [Spearman's rank-order correlation (r_s) = 0.46, $p < 10^{-7}$; Figure S8E], supporting the conclusion that K_D values measured in vitro reflect intracellular AGO binding.

When provided with K_D values for only the 12-nt *k*-mers that contained one of the six canonical sites, the biochemical+ model captured somewhat less variance (Figure 5F, green bars; $r^2 = 0.35$), and conversely when provided with K_D values for only the 12-nt *k*-mers lacking a

canonical site, the model still retained some predictive power (Figure 5F, purple bars; $r^2 = 0.06$, $p < 10^{-15}$, likelihood-ratio test). As a control, we repeated the analysis after replacing the noncanonical sites (and their K_D values) of each miRNA with those of another miRNA, performing this shuffling and reanalysis for all 309 possible shuffle permutations. When using each of these shuffled controls, performance decreased, both when considering all sites (Figure 5F, light blue bars) and when considering only the noncanonical sites (Figure 5F, pink bars), as expected if the modest improvement conferred by including noncanonical sites were due, at least in part, to miRNA pairing to those sites. This advantage of cognate over shuffled noncanonical sites was largely maintained when evaluating the results for individual miRNAs (Figure 5F). Together, our results showed that noncanonical sites can mediate intracellular repression but that their impact is dwarfed by that of canonical sites because high-affinity noncanonical sites are not highly abundant within transcript sequences. Thus, the improved performance over TargetScan achieved by the biochemical model was primarily from more accurate modeling of the effects of canonical sites.

CNN for predicting site K_D values from sequence

Our findings that binding preferences differ substantially between miRNAs and that these differences are not well predicted by existing models of RNA duplex stability in solution posed a major challenge for applying our biochemical framework to other miRNAs. Because performing AGO-RBNS for each of the known miRNAs would be impractical, we attempted to predict miRNA–target affinity from sequence using the six sets of relative K_D values and 16 miRNA-transfection datasets already in hand. Bolstered by recent successful applications of deep learning to predict complex aspects of nucleic acid biology from sequence (Alipanahi et al.,

2015; Cuperus et al., 2017; Jaganathan et al., 2019; Tunney et al., 2018), we chose a CNN for this task.

The overall model had two components. The first was a CNN that predicted relative K_D values for the binding of miRNAs to 12-nt *k*-mers (Figure S9A), and the second was the previously described biochemical model that links intracellular repression with relative K_D values (Figure 6A). The training process simultaneously tuned both the neural network weights and the parameters of the biochemical model to fit both the relative K_D values and the mRNA repression data, with the goal of building a CNN that accurately predicts the relative K_D values for all 12-nt *k*-mers of a miRNA of any sequence.

For the CNN, we chose to include only the first 10 nucleotides of the miRNA sequence, which includes the position 1 nucleotide, the seed region, and the two downstream nucleotides that could pair to a 12-nt *k*-mer. Because the *k*-mers were not long enough to include sites with 3′-supplementary pairing, we excluded the 3′ region of the miRNA. Pairs of 10-nt truncated miRNA sequences and 12-nt *k*-mers were each parameterized as a 10-by-12-by-16 matrix, with the third dimension representing the 16 possible pairs of nucleotides that could be present at each pair of positions in the miRNA and target. The first layer of the CNN was designed to learn important single-nucleotide interactions, the second layer was designed to learn dinucleotide interactions, and the third layer was designed to learn position-specific information.

The training data for the CNN consisted of over 1.5 million relative K_D values from six AGO-RBNS experiments and 68,112 mRNA expression estimates derived from 4,257 transcripts in 16 miRNA transfection experiments. Five miRNAs had data in both sets. Because some repression was attributable to the passenger strands of the transfected duplexes (Figure S9B), the model considered both strands of each transfected duplex, which allowed the neural network to learn from another 16 AGO-loaded guide sequences.

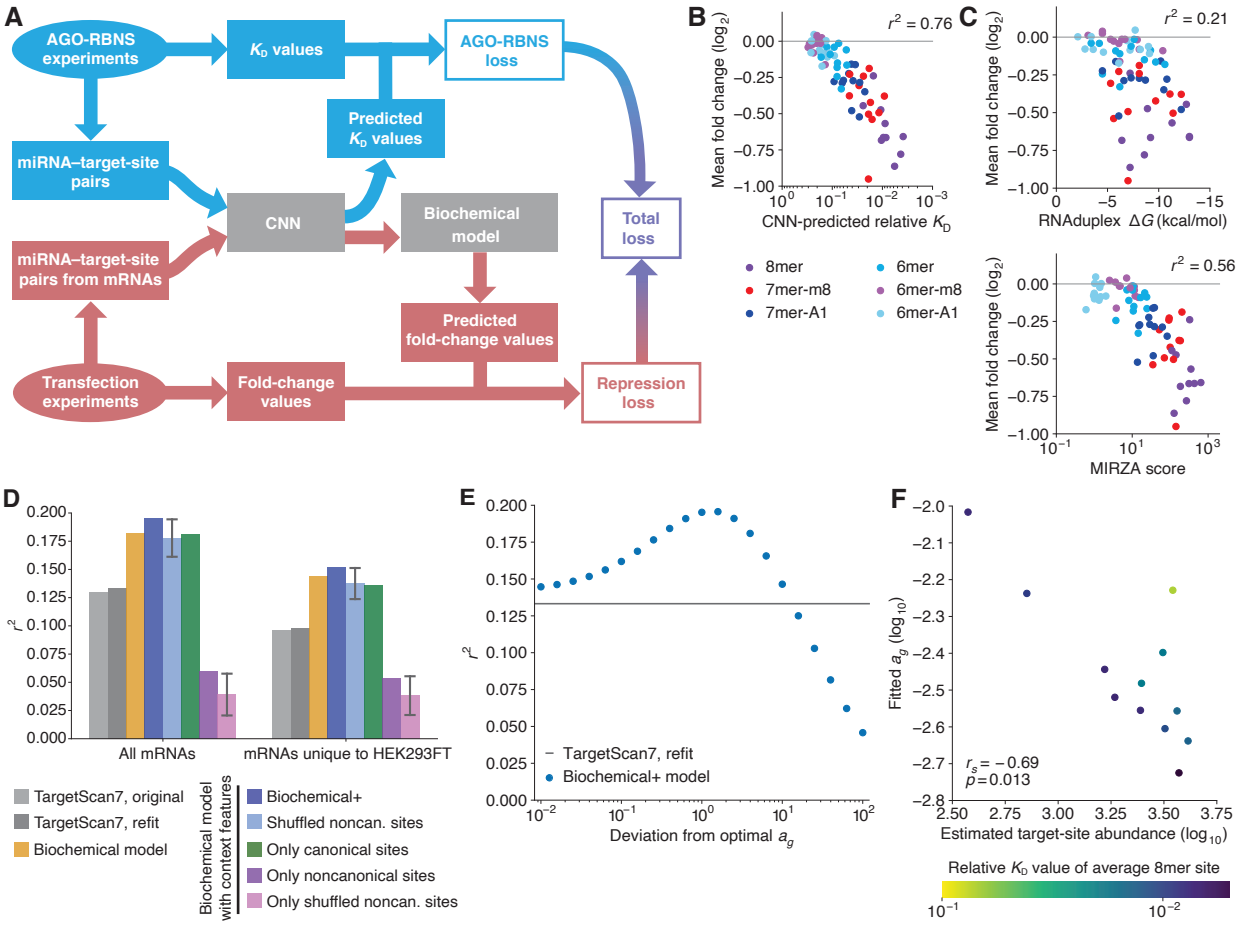


Figure 6. A CNN for predicting binding affinity from sequence.
(**A**) Schematic of overall model architecture for training on RBNS data and transfection data simultaneously. "Loss" refers to squared loss. (**B**) The relationship between repression efficacy and CNN-predicted relative K_D values for the canonical sites for the 12 test miRNAs. Everything else is the same as in Figures 3D–3I. (**C**) The relationship between repression efficacy and RNAduplex-predicted free-energy values (Lorenz et al., 2011) (top) or MIRZA scores (Khorshid et al., 2013) (bottom) for the canonical sites of the 12 test miRNAs. Everything else is the same as in (B). (**D**) Performance of the biochemical and biochemical+ models when provided the CNN-predicted relative K_D values and tested on the 12 datasets examining the effects of transfecting miRNAs into HEK293FT cells. On the left are results obtained when considering all mRNAs, and on the right are results obtained when considering mRNAs expressed in HEK293FT cells but not in HeLa cells. Everything else is the same as in Figure 5F, except shuffling results were for 250 random permutations rather than all possible permutations. (**E**) Performance of the biochemical+ model on the HEK293FT test set while allowing the a_g values to deviate from the optimal fitted values. (**F**) Relationship between fitted a_g and estimated target-site abundance (Garcia et al., 2011) for the guide strands of the 12 duplexes transfected into HEK293FT cells. Points are colored by the average relative K_D value of the 8mer site to each miRNA. The Spearman r_s and p value for the relationship are shown.

To test how well the CNN-predicted relative K_D values enabled our approach to be generalized to other miRNAs and another cell type, we generated 12 miRNA-transfection datasets in HEK293FT cells, choosing miRNAs that were not appreciably expressed in HEK293 cells (Landgraf et al., 2007) and that had not been used in any training (Figure S10). For each miRNA duplex in the test set, the CNN was used to predict relative K_D values for 12-nt *k*-mers to both the miRNA and passenger strands. As observed with the experimentally derived relative K_D values (Figures 3D–3I), substantial correspondence was observed between CNN-predicted relative K_D values for the six canonical site types of the transfected miRNAs and mean repression that these site types conferred in cells (Figures 6B and S11). This correspondence (r^2 = 0.76) substantially exceeded that observed for predictions of RNA-duplex stability in solution (Lorenz et al., 2011) and predictions derived from cross-linking results (Khorshid et al., 2013) (Figure 6C; r^2 = 0.21 and 0.56, respectively). Aside from accurately predicting the relative efficacy of sites to the same miRNA, the CNN was better able to stratify sites of the same type to different miRNAs (e.g., Figure 6B, purple dots; r^2 = 0.52, p = 0.02). Analysis of other site types suggested that the CNN had some ability to identify effective noncanonical sites for new miRNAs (Figure S11).

When the CNN-predicted K_D values and HeLa-derived global parameters were used as input for the biochemical and biochemical+ models to predict repression of individual mRNAs in HEK293FT cells, the results mirrored those observed when using relative K_D values derived from AGO-RBNS. Median (r^2 = 0.21) and overall performance (r^2 = 0.18) for the test set both exceeded those of TargetScan (r^2 = 0.12 and 0.13, respectively); overall performance improved (r^2 = 0.20) when using the biochemical+ model, implying a 50% improvement over TargetScan, and performance dropped slightly when either shuffling or omitting noncanonical sites (Figures 6D and S12A; the main exception being the results for miR-190a, for which the performance of

the biochemical+ model resembled that of TargetScan when only considering the canonical sites but substantially dropped when also considering noncanonical sites). The overall improvement over TargetScan was maintained when focusing on mRNAs that were expressed in HEK293FT cells but not HeLa cells (Figure 6D). The CNN-predicted relative K_D values also enabled the biochemical+ model to outperform TargetScan and cross-linking approaches in predicting the effects of deleting or adding a miRNA in other cellular contexts (Eichhorn et al., 2014; Lipchina et al., 2011; Zhang et al., 2018) (Figures S12B–S12D).

Although our models were improved over previous models, the highest r^2 value achieved by our models for any of our datasets was 0.37 (Figures 5F and S12A), implying that they explained only a minority of the variability in mRNA fold changes occurring upon introducing a miRNA. However, even perfect prediction of the direct effects of miRNAs was not expected to explain all of the variability; some variability was due to the secondary effects of repressing the primary targets, and some was due to experimental noise. To estimate the maximal r^2 that could be achieved by predicting the primary effects of miRNA targeting, we attempted to quantify and subtract the fraction of the fold-change variability attributable to the other two causes. For each dataset, the fraction attributable to experimental noise was estimated by examining the reproducibility between replicates in our transfection experiments, and the fraction attributable to secondary effects was inferred by assuming that primary miRNA effects only repress mRNAs, whereas secondary effects affect mRNAs in either direction (with effects distributed log normally). After accounting for these other sources of variability, the biochemical+ model provided with experimentally determined affinity values explained ~60% of the variability attributable to direct targeting (Figure S12E, median of five datasets), and when provided with CNN-predicted values it explained ~50% of the variability attributable to direct targeting (Figure S12F, median of twelve datasets).

Insights into miRNA targeting

The observation that canonical sites are not necessarily those with the highest affinity raises the question of how canonical sites are distinguished from noncanonical ones and whether making such a distinction is useful. Our results show that two criteria readily distinguished canonical sites from noncanonical ones. First, with only one exception, all six canonical site types were identified for each of the six miRNAs (the exception being the 6mer-m8 site for miR-7), whereas the noncanonical site types were typically identified for only one miRNA, and never for more than three. Second, the four highest-affinity canonical sites occupied most of the specifically bound AGO2, even for miR-124, which had the largest and highest-affinity repertoire of noncanonical sites (Figures 1F, 2, S2B, and S2C). This greater role for canonical sites was presumably because perfect pairing to the seed region is the most efficient way to bind the silencing complex; to achieve equivalent affinity, the noncanonical sites must be longer and are therefore less abundant. The ubiquitous function and more efficient binding of canonical sites explains why these site types have the greatest signal in meta-analyses of site conservation, thereby explaining why they were the first site types to be identified (Lewis et al., 2005) and justifying the continued distinction between canonical and noncanonical site types.

The potential role of pairing to miRNA nucleotides 9 and 10 has been controversial. Although some target-prediction algorithms (such as TargetScan) do not reward pairing to these nucleotides, most algorithms assume that such pairing enhances site affinity. Likewise, although one biochemical study reports that pairing to position 9 reduces site affinity (Salomon et al., 2015), another reports that it increases affinity (Becker et al., 2019). We found that extending pairing to nucleotides 9 and 10 neither enhanced nor diminished affinity in the context of seed matched sites (Figure 4), whereas extending pairing to nucleotides 9 and 10 enhanced affinity in the context of 3′-only sites (Figures 2C and 2D). These results support the idea that extensive

pairing to the miRNA 3′ region unlocks productive pairing to nucleotides 9–12, which is otherwise inaccessible (Bartel, 2018).

The biochemical parameters fit by our model provided additional insights into miRNA targeting. In the framework of our model, the fitted value of 1.8 observed for the parameter *b* suggested that a typical mRNA bound to an average of one silencing complex will experience a near tripling of its decay rate, which would lead to a ~60% reduction in its abundance. In the concentration regimes of our transfection experiments, this occupancy can be achieved with two to three median 7mer-m8 sites. In addition, our fitted value for the ORF-site penalty suggested that the translation machinery reduces site affinity by 5.5-fold.

Another parameter was a_g, that is, the intracellular concentration of AGO2 loaded with the transfected miRNA and not bound to a target site. Whereas values of the other parameters could be fit globally in HeLa cells and then used for testing, a_g was fit separately for each miRNA and passenger strand of each transfection experiment. Nonetheless, when a_g values were allowed to deviate from the fitted values, the biochemical+ model still outperformed TargetScan in predicting test-set repression over a 100-fold range of values (Figure 6E), which indicated that even with rough estimates of miRNA abundances, our modeling framework had an advantage over other predictive methods in new contexts. Information that might be used to more accurately estimate a_g values should come with the determination of these values for more miRNAs in more cellular contexts, together with the observation that, as expected (Arvey et al., 2010; Garcia et al., 2011), fitted a_g values are higher for miRNAs with lower predicted target abundance and lower general affinity for their targets (Figure 6F).

Our work replaced the correlative models of targeting efficacy with a first-principles biochemical model that explains and predicts about half of the variability attributable to the direct effects of miRNAs on their targets, raising the question of how the understanding and

prediction of miRNA-mediated repression might be further improved. Acquiring site-affinity profiles for additional miRNAs with diverse sequences will improve the CNN-predicted miRNA–mRNA affinity landscape and further flesh out the two major sources of targeting variability revealed by our study, that is, the widespread differences in site preferences observed for different miRNAs and the substantial influence of local (12-nt) site context. We suspect additional improvement will come with increased ability to predict the other major cause of targeting variability, which is the variability imparted by mRNA features more distant from the site. This variability is captured only partially by the three features added to the biochemical model to generate the biochemical+ model. Perhaps the most promising strategy for accounting for these more distal features will be an unbiased machine-learning approach that uses entire mRNA sequences to predict repression, leveraging substantially expanded repression datasets as well as site-affinity values. In this way, the complete regulatory landscape, as specified by AGO within this essential biological pathway, might ultimately be computationally reconstructed.

Methods summary

AGO2–miRNA complexes were generated by adding synthetic miRNA duplexes to lysate from cells that overexpressed recombinant AGO2, and then these complexes were purified based on affinity to the miRNA seed. RNA libraries were generated by in vitro transcription of synthetic DNA templates. For AGO-RBNS, purified AGO2–miRNA complex was incubated with a large excess of library molecules, and after reaching binding equilibrium, library molecules bound to AGO2–miRNA complex were isolated and prepared for high-throughput sequencing. Examination of *k*-mers enriched within the bound library sequences identified miRNA target sites, and relative K_D values for each of these sites were simultaneously determined by maximum

likelihood estimation, fitting to AGO-RBNS results obtained over a 100-fold range in AGO2–miRNA concentration.

Intracellular miRNA-mediated repression was measured by performing RNA-seq on HeLa cells that had been transfected with a synthetic miRNA duplex. For sites that were sufficiently abundant in endogenous 3′ UTRs, efficacy was measured on the basis of their influence on levels of endogenous mRNAs of HeLa cells. Site efficacy was also evaluated using massively parallel reporter assays, which provided information for the rare sites as well as the more abundant ones. The biochemical and biochemical+ models of miRNA-mediated repression were constructed and fit using the measured K_D values, and the repression of endogenous mRNAs was observed after transfecting miRNAs into HeLa cells. The CNN was built using TensorFlow, trained using the measured K_D values and the repression observed in the HeLa transfection experiments, and tested on the repression of endogenous mRNAs observed after transfecting miRNAs into HEK293T cells. Results were also tested on external datasets examining either intracellular binding of miRNAs by CLIP-seq or repression of endogenous mRNAs after miRNAs had been transfected, knocked down, or knocked out. The details of each of these methods are described in the section “Materials and methods.”

Acknowledgements and other information

We thank K. Heindl, T. Eisen, and T. Bepler for helpful discussions, Y. Zhou for providing processed CLIP data from miR-20a overexpression, and members of the Bartel lab for comments on this manuscript. *Funding*: This work was supported by NIH grants GM118135 (D.P.B.) and GM123719 (N.B.). D.P.B. is an investigator of the Howard Hughes Medical Institute. *Author contributions*: S.E.M. developed AGO-RBNS and associated analyses, which he implemented with help from T.P. and N.B. K.S.L. devised and implemented the biochemical model and CNN.

C.Y.S., G.M.K., and T.P. performed transfection and sequencing experiments. C.Y.S. and S.E.M. designed and performed the massively parallel reporter assay. S.E.M., K.S.L., and D.P.B. designed the study and wrote the manuscript with input from other authors. *Competing interests*: The authors declare no competing interests. *Data and materials availability*: Sequencing data are available in the Gene Expression Omnibus (accession number GSE140220), and computational tools are deposited in GitHub (https://github.com/smcgeary/agorbns and https://github.com/kslin/miRNA_models).

Materials and methods

Purification of AGO2–miRNA complexes

5′-phosphorylated RNAs of each miRNA duplex (Data S1) were synthesized (IDT), purified on a 15% polyacrylamide urea gel, and resuspended in water. A 5′-OH version of the guide strand was also synthesized (IDT) and gel purified, and 5 pmol of this RNA was 5′ radiolabeled by incubation with T4 Polynucleotide Kinase (New England Biolabs, M0201S), 2.5 μM [γ-^{32}P]-ATP (PerkinElmer, NEG035C001MC), and 1 U/μL SUPERase•In (Thermo Fisher, AM2696) at 37°C for 1 h, then passed through a P30 column (Bio-Rad, 7326250), precipitated, gel purified, and resuspended in 10 μL of annealing buffer (30 mM Tris, pH 7.5, 100 mM NaCl, 1 mM EDTA). Non-radiolabeled miRNA duplexes were generated by mixing 500 pmol of each strand, EtOH-precipitating the mixture, resuspending in 15 μL of annealing buffer, heating to near 100°C and then slow-cooling to 37°C by removing the heat block from its base. The duplex was then purified on a nondenaturing 15% polyacrylamide gel run at 8 W and 4°C for 2 h. Purified duplex was resuspended at 1 μM in annealing buffer. Radiolabeled miRNA duplexes were generated in the same way, but starting with 4 μL of radiolabeled guide strand and 20 pmol of non-radiolabeled passenger strand, heating in a 10 μL annealing reaction, and final resuspension of the sample in 10 μL of annealing buffer. The labeled duplex was treated as 50 nM, assuming a 50% loss with each gel purification.

Specific AGO–miRNA complexes were prepared using a protocol inspired by that of the Zamore lab (Flores-Jasso et al., 2013). Human embryonic kidney 293T (HEK-293T) cells were transfected with an AGO2-overexpression plasmid containing the pcDNA3.3 (Invitrogen, K8300-01) backbone driving expression from the human *AGO2* coding sequence appended with an N-terminal 3X FLAG sequence separated with a glycine-glycine-serine spacer (pcDNA3.3-3XFLAG-AGO2, Addgene plasmid #136687). Transfection was performed with Lipofectamine

2000 (Thermo Fisher, 11668019) in Opti-MEM (Thermo Fisher, 31985062), as per manufacturer instructions. After 48 h, cytoplasmic S100 extract was prepared as described (Dignam et al., 1983), except cells were lysed by passing the hypotonic suspension through a 23G needle ~10 times. The S100 extract was flash frozen in 0.5–1 mL aliquots and stored in liquid nitrogen. Stock solutions of non-radiolabeled and radiolabeled miRNA duplexes were mixed at a 10:1 ratio, and added at a 1:9 ratio to an aliquot of S100 extract to achieve final duplex concentrations of 90 and 0.45 nM, respectively. After incubation at 20°C for 2 h, 200 µL of a slurry of magnetic beads pre-bound to 500 pmol of capture oligonucleotide was added to the reaction. The magnetic-bead suspension was prepared using Dynabeads MyOne Streptavidin C1 (Invitrogen, 65002) and a biotinylated capture oligonucleotide with an 8mer site to the miRNA (Data S1) as per the manufacturer protocol, except that the beads were resuspended in equilibration buffer [18 mM HEPES, pH 7.4, 100 mM potassium acetate, 1 mM magnesium acetate, 0.01% IGEPAL® CA-630 (Sigma-Aldrich, I3021), 0.01 mg/mL yeast tRNA (Life Technologies, 15401011), and 0.1 mg/mL BSA (New England Biolabs, B9000S)]. After incubation at 20°C for 30 min, the beads were washed five times with 200 µL of equilibration buffer, and then five times with 200 µL of equilibration buffer supplemented with 2 M potassium acetate. The sample was eluted by incubating for 2 h with 10 µM competitor oligonucleotide (Data S1), which was complementary to the capture oligo, in 100 µL of equilibration buffer supplemented with 1 M potassium acetate. Tagged AGO2 was then further purified using 20 µL of Anti-FLAG M2 magnetic beads (Sigma-Aldrich, M8823), as per the manufacturer protocol but using equilibration buffer rather than the buffer suggested by the manufacturer. The AGO2–miRNA complex was eluted from the Anti-FLAG beads by incubating with 60 µL of equilibration buffer containing 146 ng/µL 3X FLAG peptide (Sigma-Aldrich, F4799) at 22°C and shaking at 1300 rpm for 1 h. DTT and glycerol were each added to the eluate to reach the final concentration of the protein storage buffer [13

mM HEPES, pH 7.4, 72 mM potassium acetate, 0.72 mM magnesium acetate, 2.2 mM Tris-HCl, pH 7.4, 4.3 mM NaCl, 0.0072% (v./v.) IGEPAL CA-630, 0.0072 mg/mL yeast tRNA, 0.072 mg/mL BSA, 5 mM DTT, and 20% (v./v.) glycerol]. The stock concentration of each purified AGO2‑miRNA complex ranged from 0.42‑1.1 nM, as estimated by autoradiography of 1 µL of the sample spotted onto a Hybond nylon (Thermo Fisher, 45001147) filter membrane alongside 1 µL of the initial S100 extract loaded with ~90 nM miRNA duplex.

Three independent preparations of AGO2–miR-1 were made. The first and second were used to determine the consistency of AGO-RBNS results (Figure S1B); the second was used for de novo site identification and all other analyses performed, and the third was used as a replicate for *de novo* site identification (see "De novo site identification"). Two independent preparations of AGO2–miR-124 and AGO2–miR-7 were also made, with the first prepared as described above and the second prepared with the following changes: 1) S100 extracts were prepared from HEK293FT cells rather than HEK293T cells, 2) cells were harvested 24 h after transfection, 3) miRNA duplexes were not gel purified prior to transfection, 4) AGO2–miR-124 was eluted from the capture oligo–bead slurry with 7.5 µM competitor oligo in 100 µL of equilibration buffer, and 5) AGO2–miR-7 was incubated with a slurry of magnetic beads pre-bound to 50 pmol of capture oligonucleotide and subsequently eluted from the capture oligo–bead slurry with 0.75 µM competitor oligonucleotide in 100 µL of equilibration buffer. These second preparations each had substantially reduced residual competitor oligo and were used as replicates for de novo site identification, which helped prevent sites from being identified by virtue of complementarity to the competitor oligo (see "De novo site identification").

Small-RNA sequencing of AGO–miRNA preparations

Purified AGO2–miR-1 and purified AGO2–miR-155 were each extracted with TRI Reagent (Sigma-Aldrich, T9424), and before separating aqueous and organic phases, two non-human miRNAs (dme-miR-14-5p and xtr-miR-427, Data S1) were added for inter-library comparison, and radiolabeled 18- and 30-nt standards (Data S1) were added for size selection. After gel purification on a 15% polyacrylamide urea gel, RNA was ligated to a pre-adenylated 3′ adapter (Data S1) using T4 RNA Ligase 2, truncated KQ (New England Biolabs, M0373S) in a reaction supplemented with 10% (v./v.) PEG 8000 (Sigma-Aldrich, 25322-68-3). After gel purification on a 10% polyacrylamide urea gel, RNA was ligated to a 5′ adapter (Data S1) using T4 RNA Ligase I (New England Biolabs, M0204) in a reaction supplemented with 10% (v./v.) PEG 8000. To reduce ligation biases, this adapter had 14 random-sequence nucleotides at its 3′ end. After gel purification on an 8% polyacrylamide urea gel, RNA was reverse transcribed with SuperScript II (Thermo Fisher, 18064014), and the cDNA was amplified for 8–12 cycles with Phusion (New England Biolabs, M0530) DNA polymerase. Amplified DNA was purified on an 8% polyacrylamide, 90% formamide gel and submitted for sequencing. A step-by-step protocol for constructing libraries for small-RNA sequencing is available at http://bartellab.wi.mit.edu/protocols.html. Libraries were sequenced on an Illumina HiSeq 2500 with 40-nt single-end reads. To count the miRNAs in each library, reads were first subjected to quality-control filtering (see "RBNS read quality control," steps 1–5), and then the 14 nt of random adaptor sequence at the 5′ end and the constant adaptor sequence at the 3′ end were removed. Reads greater than 18 nt in length after adaptor trimming were mapped by querying the first 18 nt of each against a list of the first 18 nt of human miRNAs annotated in miRbase v22.1, supplemented with the 5′ and 3′ adapter sequences, the 18- and 30-nt marker sequences, and the dme-miR-14-5p and xtr-miR-427 sequences. Counts were normalized to the total number of

counts corresponding to human miRNAs to obtain the counts-per-million (cpm) values reported in Figure S1A.

Preparation of RNA libraries for AGO-RBNS

Four libraries of DNA oligonucleotides, each containing a central region of 37 random-sequence positions (Data S1), were synthesized (IDT) and purified on 6% polyacrylamide urea gels. Each RNA library was then generated from a 500 µL in vitro transcription reaction using T7 RNA polymerase (Rio, 2013), 1 µM gel-purified template DNA, 1 µM T7 forward primer (Data S1), 8 mM GTP, 5 mM CTP, 5 mM ATP, 2 mM UTP, 5 mM DTT, 40 mM Tris-HCl, pH 7.9, 2.5 mM Spermidine, 26 mM $MgCl_2$, and 0.01% (v./v.) Triton X-100, at 37°C for 2.5 h. The reaction was then incubated with 10 µL of TURBO DNase (Thermo Fisher, AM2238) at 37°C for 10 min, and then the RNA purified on a 6% polyacrylamide urea gel. 200 pmol of library was then 5′-cap labeled with Vaccinia Capping System (New England Biolabs, M2080S) in a reaction containing 0.1 mM GTP and 3.33 µM [α-32P]-GTP (PerkinElmer, BLU006H250UC), according to the manufacturer's protocol. The sample was then extracted with phenol–chloroform, precipitated, resuspended in 5 µL of H_2O, dephosphorylated using Calf Intestinal Phosphatase (CIP, New England Biolabs, M0290S) at 37°C for 45 min according to the manufacturer's protocol, and then gel purified.

Preparation of AGO-RBNS quantification standards

Defined RNAs were added to each AGO-RBNS sequencing library at the step of the Proteinase K incubation (see "AGO-RBNS") to enable quantitative comparison of the RNA recovered in each binding sample. These quantification standards (Data S1) were generated by in vitro transcription of the corresponding PCR templates (Data S1), followed by TURBO DNase

treatment, gel purification, CIP treatment, and gel purification, as described for the RNA libraries (see "Preparation of RNA libraries for RBNS").

AGO-RBNS

Each AGO-RBNS experiment included five binding reactions that spanned a 100-fold concentration range of AGO–miRNA complex. For each experiment, the greatest concentration was that in which the stock solution of the complex comprised 40% (v./v.) of the binding reaction, and for each of the four additional reactions in each series, this stock was serially diluted 3.16-fold into protein storage buffer, resulting in the 100-fold range of the complex over five reactions. Each experiment also included a mock binding reaction using protein storage buffer without AGO–miRNA complex. Each binding reaction was performed in 10 µL, and in addition to the AGO–miRNA complex, each reaction contained 100 nM RNA library (see "Preparation of RNA libraries for RBNS"), 16 mM HEPES, pH 7.4, 89 mM potassium acetate, 0.89 mM magnesium acetate, 0.043 ng/µL 3X FLAG peptide, 0.87 mM Tris-HCl, pH 7.5, 1.7 mM NaCl, 0.0029% IGEPAL CA-630, 0.0089 mg/mL yeast tRNA, 0.029 mg/mL BSA, 7 mM DTT, 1 U/µL SUPERase•In, and 8% (v./v.) glycerol. Reactions were incubated for 2 h at 37°C and then filtered through stacked Protran nitrocellulose (Sigma-Aldrich, Z670898) and Hybond nylon filter membranes. To ensure constant temperature throughout the procedure, incubations and filtering were performed in a 37°C constant-temperature room, using supplies that had been pre-equilibrated to 37°C. Filtering was through circular membranes (0.5-inch diameter) that had been punched from stock, pre-equilibrated with filter-binding buffer (18 mM HEPES, pH 7.4, 100 mM potassium acetate, and 1 mM magnesium acetate), stacked with the nitrocellulose membrane atop the nylon membrane onto the internal pedestal of a Whatman filter holder (Sigma-Aldrich, WHA420100) that was inserted into a closed valve of a Visiprep vacuum

manifold (Sigma-Aldrich, 57250-U). For filter binding, 100 µL of filter-binding buffer was applied to the top filter, the valve was opened, the binding reaction was applied, and the membrane stack was immediately washed with 100 µL of ice-cold wash buffer (filter-binding buffer supplemented with 5 mM DTT). The two membranes were then separated and allowed to air-dry. After phosphorimaging to monitor binding, the nitrocellulose membranes were each incubated with 1 µg/µL Proteinase K (Life Technologies, 25530049) in 400 µL of Proteinase K buffer (50 mM Tris-HCl, pH 7.4, 50 mM NaCl, and 10 mM EDTA). A Proteinase K reaction was also prepared with 1.5 pmol of the 5′ cap-labeled input library. Quantification standards (Data S1) were added to each reaction at an expected ratio of 1:1000, allowing for quantitation of RNA recovery. After 10 min at 37°C, SDS was added at 0.5% (w./v.) final concentration, and reactions were incubated at 65°C for 45 min with shaking on a thermomixer. Samples were then phenol–chloroform extracted, EtOH-precipitated, resuspended in 5 µL of water, and reverse transcribed in a 30 µL reaction using SuperScript II (removing 3 µL prior to addition of enzyme as an "RT-minus" control). RNA was degraded by adding 5 and 0.5 µL of 1 M NaOH to the RT-plus and RT-minus reactions, respectively, and incubating at 90°C for 10 min. The reactions were then neutralized by adding 25 and 2.5 µL of 1 M HEPES, pH 7.0, to the RT-plus and RT-minus reactions, respectively. Each reaction was then brought to 60 µL with water and passed through a P30 column, and then 4 µL of each reaction was amplified in a 50 µL reaction with Phusion. Both the RT-plus and RT-minus–derived reactions were run on an 8% polyacrylamide, 90% formamide gel, and the RT-plus–derived amplicons were purified and then sequenced on an Illumina HiSeq 2500 with 40-nt single-end reads.

miRNA transfections and mRNA-seq library preparation

RNAs of each miRNA duplex (Data S1) were synthesized (IDT), resuspended at 200 μM in IDT Duplex Buffer (30 mM HEPES, pH 7.5, and 100 mM potassium acetate), annealed as described above, and transfected without gel purification. For each transfection of HeLa and HEK293FT cells, 2.5 and 2.1 million cells, respectively, were plated in a 10 cm dish supplied with 10 mL of media (DMEM + 10% FBS). After 24 h of culture, the cells were supplied with fresh media and transfected with 1 nmol of RNA duplex using Lipofectamine RNAiMAX (Thermo Fisher, 13778150) and Opti-MEM (Thermo Fisher, 31985062) as per the manufacturer's protocol modified to achieve a final duplex concentration of 100 nM. After 24 h, cells were harvested, and total RNA was extracted using TRI Reagent (Sigma-Aldrich, T9424) according to the manufacturer's protocol. RNA-seq libraries were prepared from 10 μg of total RNA per sample using the Bioo Nextflex Directional Rapid RNA-seq kit with poly(A)-selection beads (PerkinElmer, #NOVA-5138-07). Transfection and library preparation were performed in replicates, with the two replicates of each miRNA duplex performed in different batches, performing a total of five batches for the HeLa transfections and three batches for the HEK293FT transfections. Sequencing was on an Illumina HiSeq 2500 with 40-nt single-end reads for the HeLa transfections, and 50-nt single-end reads for the HEK293FT transfections.

Massively parallel reporter library

A reporter-plasmid library was designed to assay the efficacy of all 163 miRNA sites originally identified in the initial AGO-RBNS replicates of this study (McGeary et al., 2018), each within many different sequence contexts. Each library member was designed to express (from the pEF1a promoter) a *GFP* mRNA with a 146-nt variable-sequence region spanning positions 34–179 of its 306-nt 3′ UTR. Each variable-sequence region harbored a single miRNA site centered

either at position 106 or between positions 106 and 107, depending on whether the site was of odd or even length. The remaining positions of each variable-sequence region were chosen by weighted sampling of dinucleotides according to the average frequency of each over all human 3′ UTR sequences, while excluding any additional site to any of the six miRNAs. Each of the 163 sites was designed to be presented in 184 contexts, yielding 29,993 UTR possibilities (data S3). The parental plasmid was based on pCMV-GFP (Addgene, plasmid #11153), but with positions 4405–4479 and 1–580 (a 655-bp contiguous segment spanning the ends of the deposited plasmid map) replaced with positions 2632–3792 of pJA291 (Addgene, plasmid #74487) and positions 1335–1339 replaced with a 16-nt sequence containing a BstXI site (ATAACCACGCTGATGG), with positions 1669–2842 of eSpCas9(1.1) (Addgene, plasmid #71814) immediately downstream. The first modification conferred the eGFP pre-mRNA with an intron so as to better resemble endogenous genes, and also replaced the CMV promoter with an EF1-alpha promoter. The second modification removed the 5′ splice site consensus sequence overlapping the STOP codon, and introduced two BstXI sites separated by 1229 nucleotides into the 3′ UTR. The DNA library of variable-region sequences (Twist Biosciences, Oligo Pools order, Data S3) was amplified with primers adding 1) homology to the 5′ PCR primer used for small RNA-seq library preparation, and 2) homology to each of the BstXI sites at the very 5′ and 3′ ends of the amplicon (Data S1). This amplicon was incubated with the large fragment from a BstXI digest of the parental plasmid in a Gibson assembly reaction (New England Biolabs, E2611S) to produce the reporter-plasmid library. The Gibson reaction was electroporated into OneShot Top10 Electrocomp *E. coli* (Thermo Fisher, C404050), and bacteria from all ten electroporations were plated onto 66 10 cm LB agar plates. After 16 h of bacterial growth under ampicillin selection, bacteria were harvested, and the reporter-plasmid library was purified by MAXI-prep (Qiagen, 12362).

Massively parallel reporter assay

Each massively parallel reporter assay was performed first by plating 0.724 million HeLa cells in a 10 cm dish supplied with 10 mL media (DMEM + 10% FBS). After 24 h of culture, the cells were supplied with fresh media and transfected with one of the six miRNA duplexes or a mock using Lipofectamine RNAiMAX as per the manufacturer's protocol modified to achieve a final duplex concentration of 144 nM (or 0 nM in the case of the mock). After 24 h of culture, the cells were supplied with fresh media and transfected with 5.8 µg of reporter library diluted in 28.9 µg of pUC19 carrier plasmid using Lipofectamine 2000 (Thermo Fisher, 11668019) as per the manufacturer's protocol. After 24 h, cells were harvested by decanting the media, washing and decanting twice with ice-cold PBS, and then adding 362 µL of lysis buffer [10 mM Tris-HCl, pH 7.4, 5 mM $MgCl_2$, 100 mM KCl, 1% (v./v.) Triton X-100, 2 mM DTT, 0.02 U/µL SUPERase•In, and 1 tablet per 10 mL cOmplete EDTA-free Protease Inhibitor] evenly over the surface of the plate. Cells were then scraped off the plate and transferred to a 1.5 mL microcentrifuge tube, and lysed by gently passing the cell suspension through a 26G needle four times. The lysed cells were then pelleted at 1300 × *g* for 10 min, and the supernatants (~450 µL) each transferred to a new tube. Total RNA was extracted by first splitting each sample into three separate aliquots (~150 µL each) and adding 1 mL of TRI Reagent to each aliquot and pooling the extracted RNA. Half of the recovered RNA from each sample was then treated with TURBO DNase, using 1 µL of enzyme in 50 µL of total reaction volume per 10 µg of total RNA, incubating at 37°C for 30 min. The samples were then re-extracted with phenol–chloroform, EtOH-precipitated, and resuspended in water to their original volumes. Reverse transcription, PCR, and formamide gel purification to generate amplicons for RNA-seq were performed as described (see "AGO-RBNS") with the following modifications: 1) the RT primer was designed

to reverse transcribe the variable 3′-UTR region of the reporter library and add homology to the 3′ PCR primer used for small RNA-seq library preparation (Data S1), 2) the volumes of the RT reactions were scaled up, using 1 µL of SuperScript II in 30 µL of total reaction per 5 µg of total RNA, 3) after base-hydrolysis of the RT reactions and neutralization with HEPES, each RT reaction was EtOH-precipitated and resuspended in 60 µL of water before the P30 step, and 4) after performing a pilot PCR using 4 µL of the cDNA in a 50 µL reaction to determine the minimal number of cycles to achieve amplification, the remaining 56 µL of cDNA was amplified in seven 100 µL PCR reactions. These seven reactions were combined, and DNA was precipitated and resuspended for formamide-gel purification. These modifications, which scaled up the input and the amplification volume, were designed to increase the number of distinct library mRNAs contributing to the measured expression of each variant. All seven conditions (the six miRNA duplex transfections and the mock transfection) were performed in duplicate, and the fourteen samples were sequenced with multiplexing on two lanes of an Illumina HiSeq 2500 run in rapid mode with 100-nt single-end reads. For analysis, reads were first subjected to quality-control filtering (see "RBNS read quality control," steps 1–5). Reads passing these criteria were then assigned to one of the 29,992 sequences designed for the library, requiring a perfect match to the sequence. For each sequence, counts were normalized to the total number of perfectly matching counts to obtain counts per million (cpm).

RBNS read quality control

Each RBNS sequencing read was used if it satisfied the following criteria: 1) it passed the Illumina chastity filter, as indicated by the presence of the number 1 rather than 0 in the final position of the fastq header line, 2) it did not contain any "N" base calls, 3) it did not contain any positions with a Phred quality score (*Q*) of B or lower, 4) the sequenced 6-nt sample-

multiplexing barcode associated with the read was identical to one of the barcodes used when generating the sequencing library, 5) it did not match either strand of the phi-X genome, 6) it did not nearly match (allowing up to two single-nucleotide-substitutions/insertion/deletions) the standards added to the samples during library workup, and 7) it contained either a TCG at positions 38–40 in the library of the first AGO2–miR-1 experiment or a TGT at these positions for all other experiments.

De novo site identification

To identify sites of an AGO–miRNA complex using RBNS results, we performed an analysis in which we 1) calculated the enrichment of all 10-nt *k*-mers in the library from the binding reaction with the greatest concentration of AGO–miRNA, 2) defined a site by computationally assisted manual curation of the ten most highly enriched 10-nt *k*-mers, as outlined below, and 3) removed all reads containing the identified site from both the input and the bound libraries corresponding to that AGO-RBNS experiment. This three-step process was repeated until no 10-nt *k*-mer with an enrichment >10-fold remained. For miR-1, miR-124, and miR-7, this process was performed with two separate AGO-RBNS experiments, each of which had used a separately purified AGO–miRNA complex (see "Purification of AGO2–miRNA complexes"). The AGO-RBNS experiments performed with second purifications of AGO2–miR124 and AGO2–miR-7 included technical replicate samples that were sequenced independently, with the reads combined for these analyses.

To identify a miRNA site at each iteration, we queried each of the ten most highly enriched *k*-mers for its extent of complementarity to the miRNA. This was performed by first testing for perfect complementarity to 10 contiguous positions of the miRNA. In the case of imperfect complementarity, the *k*-mer was further tested for any of the following: 1)

complementarity to nine contiguous miRNA positions, allowing a single internal bulged target nucleotide, 2) complete complementarity to the miRNA at all ten positions while allowing for wobble pairing, 3) complementarity to the miRNA at nine positions of the 10-nt *k*-mer with an internal non-wobble mismatch position, 4) complementarity to the miRNA at nine positions of the 10-nt *k*-mer, while allowing wobble pairing and a single bulged target nucleotide, or 5) complementarity to the miRNA at eight positions within the 10-nt *k*-mer, allowing both a bulged nucleotide and an internal mismatch position. *k*-mers with miRNA complementarity starting between miRNA positions 1–5 and ending beyond position 8 were defined as ending at position 8, to prevent falsely characterizing flanking nucleotide content at positions 9 and 10 as a preference for complementarity to miRNAs with an A or a U at these positions. Any identified pairing configurations without full Watson–Crick complementarity were stored, and then the process was repeated on the two 9-nt sub-*k*-mers within the 10-nt *k*-mer, the three 8-nt sub-*k*-mers within the 10-nt *k*-mer, etc., until a sub-*k*-mer was identified as having full Watson–Crick complementarity to a region of the miRNA.

The list of candidate sites identified for a 10-nt *k*-mer were then ranked using a scoring system that rewarded 1) each Watson–Crick pair within the site (preferentially to nucleotides 2–8, 12–16, 17–22 or 23, and 9–11, in that order), 2) each dinucleotide of Watson–Crick pairing (uniformly across the miRNA sequence), 3) contiguous pairing to miRNA nucleotides 2–5, and 4) A/U content external to the sub-*k*-mer classified as participating in the miRNA–target interaction, and penalized 1) bulged nucleotides, 2) wobble pairs, 3) mismatched pairs, and 4) G content outside of the internal region of the 10-nt *k*-mer defined as participating in the miRNA–target interaction. The weights associated with each reward and penalty were tuned such that the site identified within each 10-nt *k*-mer was consistent with that identified by visual inspection, with the rationale that correctly identified sites <10 nt in length would be present in more than

one of the ten most enriched 10-nt *k*-mers—each instance in a different flanking context, with a preference for A and U nucleotides within this flanking sequence. The script (with tuned weights) used to score candidate sites is available at https://github.com/smcgeary/agorbns. This inherently ad hoc approach was used to evaluate sites in a consistent manner for all miRNAs, thereby mitigating two major sources of ambiguity when identifying miRNA sites: 1) the variable extent of sequence redundancy within miRNAs (e.g., miR-1: UGGAAUGUAAAGAAGUAUGUAU, let-7a: UGAGGUAGUAGGUUGUAUAGGU), and 2) the potential for conflating favorable site context with extended pairing when analyzing A/U-rich miRNAs [e.g., the choice of designating AUAAUUCCA as a miR-1 8mer-w7bA(6.7) site or as an instance of a 6mer-A1 site (AUUCCA) in a favorable flanking nucleotide context (AUA)].

If the most enriched 10-nt *k*-mer paired (allowing wobbles) throughout its length to the 3′ end of the miRNA sequence, enrichment of all 11-nt *k*-mers was also calculated, and if the most highly enriched 11-nt *k*-mer containing the 10-nt *k*-mer also fully paired to the miRNA, the site was designated as an 11-nt site. Likewise, if the site ascribed to the most enriched 10-nt *k*-mer was a 7mer-m8-like site with flanking A/U nucleotides only in the 5′ region of the *k*-mer and if the nucleotide at miRNA position 2 paired to the 10th position of the *k*-mer (and if the 8mer-like version of the site hadn't yet been identified), the enrichment of 11-nt *k*-mers was calculated, and the site type was designated as the 8mer-like form if the most highly enriched 11-nt *k*-mer containing the 7mer-m8-like site included an A at target position 1.

When identifying sites with no obvious pairing to the miRNA (i.e., ≤4 nt of pairing, including wobble pairing, or 5 nt of pairing but with non-A/U-rich sequences flanking the proposed segment of pairing), the top 9-nt sub-*k*-mer was preliminarily assigned as the site. In the case of miR-1, miR-124 and miR-7, for which the de novo site identification was performed independently for two AGO-RBNS replicates (see "Purification of AGO2–miRNA complexes"),

a 9-nt *k*-mer was retained only if a similar *k*-mer was identified in the other replicate. In the cases of let-7, miR-155, and lsy-6, for which only one AGO-RBNS experiment was performed, sites with no obvious pairing to the miRNA were not retained if they had ≥6 contiguous pairs to the competitor oligo used for purification of the AGO–miRNA complex. The 9-nt *k*-mers still under consideration included the CGCUUCCGC motif for miR-1, the UGCACUUUA, AGCACUUUA, and CGCACUUUA motifs for let-7a, the AACGAGGAA, UAACGAGGA, AACGAGGAU, AACGAGGAG, and AUAACGAGG motifs for miR-155, the AACGAGGAA motif for lsy-6, and the CGCUUCCGC, CUUCCGCUG, and GCUUCCGUU motifs for miR-7. Owing to the apparent similarity of these 9-nt *k*-mers for each miRNA, the representative site was chosen to be the most enriched 8-nt sub-*k*-mer contained within one of the 9-nt *k*-mers listed here, determined at the first iteration of site removal for which one of these 9-nt *k*-mers was found within the top 10-nt *k*-mer. These were the GCUUCCGC motif for miR-1, the GCACUUUA motif for let-7a, the AACGAGGA motif for miR-155, the AACGAGGA motif for lsy-6, and the GCUUCCGC motif for miR-7.

We note that our requirement of a >10-fold enrichment of 10-nt *k*-mers did not necessarily yield sites with K_D values >10-fold better than the no-site value. For example, the miR-1 6mer-m8 site was identified through this procedure, despite its K_D value being only 3.5-fold better than the no-site value (Figure 1F). This site was identified because some 10-nt *k*-mers with the 6mer-m8 site had the site within a favorable sequence context (e.g., with A/U-rich dinucleotides flanking both sides of the site), and these *k*-mers that presented the site in a favorable context were enriched >10-fold. With our protocol, the shorter sites had more opportunity to benefit from favorable flanking nucleotides than did the longer sites.

The procedure for identifying sites was modified for miR-124, for which various sites with imperfect pairing to the seed (due to internal bulges, wobble pairing, or mismatched

nucleotides) had unusually high binding affinity when preceded by an AA 5′-flanking dinucleotide. Because the effect of this 5′ flanking dinucleotide was substantially greater than the general flanking-dinucleotide effect (Figures 4 and S6), only for these sites, and only for miR-124, they are reported as AA-[site type] to distinguish them from the generic benefit of A/U-rich flanking dinucleotides (Figure 2C).

Determination of K_D values from AGO-RBNS data

Overview of maximum likelihood estimation–based approach

Relative K_D values for a set of sites were simultaneously determined by maximum likelihood estimation (MLE). In this statistical method, the parameter values $\boldsymbol{\theta}$ of a mathematical model are fit to maximize the log-likelihood function

$$\ln \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \ln p(\boldsymbol{y} \mid \boldsymbol{x}(\boldsymbol{\theta})), \tag{2.1}$$

where $p(\boldsymbol{y} \mid \boldsymbol{x}(\boldsymbol{\theta}))$ is the probability of observing the sequencing counts $\boldsymbol{y}$ given the model-simulated abundances $\boldsymbol{x}(\boldsymbol{\theta})$ (itself a function of $\boldsymbol{\theta}$). We first describe the derivation of $\boldsymbol{x}(\boldsymbol{\theta})$ and then of $f_{cost}(\boldsymbol{x})$, a cost function scaling monotonically with $\ln p(\boldsymbol{y} \mid \boldsymbol{x}(\boldsymbol{\theta}))$ and therefore having a minimum value coincident with the MLE parameter estimates. We then derive the gradient of the cost function

$$f_{grad}(\boldsymbol{\theta}) = \nabla f_{cost}(\boldsymbol{x}(\boldsymbol{\theta})). \tag{2.2}$$

The optimization routine was performed with the *optim* function in R (R Core Team, 2014) using the L-BFGS-B method, supplying both $f_{\text{cost}}(\boldsymbol{x})$ and $f_{\text{grad}}(\boldsymbol{x})$ to the optimizing function as compiled C scripts through the *.C* interface. This enabled efficient, simultaneous estimation of a large set (>50,000) of K_D values per AGO-RBNS experiment.

Derivation of $\boldsymbol{x}(\boldsymbol{\theta})$

The function $\boldsymbol{x}(\boldsymbol{\theta})$ produces an $m \times n$ matrix where each element x_{ij} specifies a model estimate of the concentration of library RNA molecules of site type i recovered from binding reaction j for a particular AGO-RBNS experiment. The dimensions m and n are therefore determined by the number of distinct types of sites (where library RNA molecules that do not contain a site constitute the mth site type) and the total number of binding reactions comprising that AGO-RBNS experiment, respectively. In practice, $n = 5$ for all experiments other than that with AGO2–miR-7, for which $n = 4$ because the 4% dilution sample was discarded for technical reasons. This calculation requires as input the total concentration of each site type $\boldsymbol{l} = (l_1, \ldots, l_m)$, the total concentration of AGO–miRNA complex (hereafter referred to as "AGO") in each binding reaction $\boldsymbol{a} = (a_1, \ldots, a_n)$, the K_D value describing the binding between AGO and each site type $\boldsymbol{K} = (K_1, \ldots, K_m)$, and the concentration of library RNA recovered due to nonspecific binding to the nitrocellulose filter b, which is assumed to be constant across all five samples and therefore given by a single parameter. The vector $\boldsymbol{l}$ is estimated using

$$\boldsymbol{l} = \frac{\boldsymbol{y}^l}{\sum_{i=1}^{m} y_i^l} \times 100 \text{ nM}, \tag{2.3}$$

where $\boldsymbol{y}^l$ is the vector of read counts corresponding to each site type as measured in the sequencing of the input library. Each element a_j of $\boldsymbol{a}$ is calculated from the experimentally determined dilution series

$$\begin{aligned} \boldsymbol{a} &= a \times \boldsymbol{s} \\ &= a \times (0.4\%,\ 1.27\%,\ 4\%,\ 12.7\%,\ 40\%), \end{aligned} \tag{2.4}$$

where a is the stock (pre-dilution) concentration of AGO, and so only the parameter a is included in $\boldsymbol{\theta}$. The set of parameters to be optimized is therefore

$$(K_1, K_2, \ldots, K_m, a, b). \tag{2.5}$$

Because these parameters represent either binding affinities or concentrations, for which negative values are physically meaningless, $\boldsymbol{x}(\boldsymbol{\theta})$ performs an exponential transformation on $\boldsymbol{\theta}$:

$$\begin{aligned} K_1 &= e^{\theta_1} \\ &\vdots \\ K_m &= e^{\theta_m} \\ a &= e^{\theta_{m+1}} \\ b &= e^{\theta_{m+2}}, \end{aligned} \tag{2.6}$$

such that any negative parameter values queried during the optimization routine will correspond to a value between 0 and 1 within the biochemical equations of $\boldsymbol{x}(\boldsymbol{\theta})$.

The recovered concentration of site type i in sample j is given by

$$x_{ij} = c_{ij} + g_{ij}, \tag{2.7}$$

where c_{ij} and g_{ij} are the concentration of AGO-bound and nonspecifically recovered forms of the site type, respectively. The nonspecifically recovered RNA g_{ij} is assumed to only come from the unbound sites in the binding reaction, such that

$$g_{ij} = \alpha_j l_{ij}^f, \tag{2.8}$$

where l_{ij}^f represents the concentration of the unbound form of site type i in sample j, and α_j is a sample-specific proportionality constant. Making the assumption that the total concentration of nonspecifically recovered RNA (summed over all m site types) is equal to b (= $e^{\theta_{m+2}}$), yields

$$\sum_{i=1}^{m} g_{ij} = b$$

$$\sum_{i=1}^{m} \alpha_j l_{ij}^f = b$$

$$\alpha_j = \frac{b}{\sum_{i=1}^{m} l_{ij}^f}. \tag{2.9}$$

Substituting for α_j in equation (2.8) using equation (2.9), and further substituting for g_{ij} in equation (2.7) yields

$$x_{ij} = c_{ij} + \frac{b}{\sum_{i'=1}^{m} l_{i'j}^f} l_{ij}^f. \tag{2.10}$$

By invoking the conservation of mass for each site type (i.e., $c_{ij} + l_{ij}^f = l_i$), equation (2.10) can be expressed as

$$x_{ij} = c_{ij} + b\frac{l_i - c_{ij}}{\sum_{i'=1}^{m}\left(l_{i'} - c_{i'j}\right)}$$

$$x_{ij} = c_{ij}\left(1 - \frac{b}{L - C_j}\right) + l_i \frac{b}{L - C_j}, \tag{2.11}$$

where $L = \sum_{i=1}^{m} l_i$ represents the total concentration of the RNA library in the reaction (experimentally set to 100 nM), and $C_j = \sum_{i=1}^{m} c_{ij}$ represents the total concentration of bound RNA library in sample *j*.

Equation (2.11) gives the model-predicted values x_{ij} in terms of only known quantities (l_i, its sum *L*, and *b*), and the concentration of bound form of each site type c_{ij}. This quantity can

be expressed as a function of the K_i ($=e^{\theta_i}$ where $i \in [1 .. m]$) parameter values by invoking the definition of K_D:

$$K_i \equiv \frac{a_j^f l_{ij}^f}{c_{ij}}, \tag{2.12}$$

where a_j^f represents the concentration of unbound AGO in sample j. As before, l_{ij}^f is substituted by invoking the conservation of mass, yielding

$$K_i = \frac{a_j^f (l_i - c_{ij})}{c_{ij}}, \tag{2.13}$$

which is rearranged to give

$$c_{ij} = \frac{l_i a_j^f}{a_j^f + K_i}. \tag{2.14}$$

Using equation (2.14) to substitute for c_{ij} in equation (2.11) yields

$$x_{ij} = l_i \left(\frac{a_j^f}{a_j^f + K_i} \left(1 - \frac{b}{L - C_j} \right) + \frac{b}{L - C_j} \right), \tag{2.15}$$

and since $C_j = \sum_{i=1}^{m} c_{ij}$,

$$x_{ij} = l_i \left(\frac{a_j^f}{a_j^f + K_i} \left(1 - \frac{b}{L - \sum_{i'=1}^{m} \frac{l_{i'} a_j^f}{a_j^f + K_{i'}}} \right) + \frac{b}{L - \sum_{i'=1}^{m} \frac{l_{i'} a_j^f}{a_j^f + K_{i'}}} \right). \tag{2.16}$$

This is the final form of the function, wherein read abundances are modeled from the fixed vector $\boldsymbol{l}$ (and its sum L) and the parameter vector $\boldsymbol{\theta}$ where $K_i = e^{\theta_i}$ for $i \in [1 .. m]$, $a = e^{\theta_{m+1}}$, and and $b = e^{\theta_{m+2}}$, and whose values are iteratively updated during the optimization routine. Equation

(2.16) cannot be used directly; it requires a value for the concentration of unbound AGO in sample *j*, a_j^f. This value is obtained by invoking the conservation of mass for AGO in sample *j*:

$$a_j = a_j^f + \sum_{i=1}^{m} c_{ij}. \tag{2.17}$$

Because each c_{ij} value is itself a function of $\boldsymbol{l}$, $\boldsymbol{K}$, and a according to equation (2.14), equation (2.17) specifies a single value of a_j^f. However, this equation cannot be rearranged to an explicit expression for a_j^f. Therefore, each time $\boldsymbol{x}$ is calculated during the optimization routine requires that a_j^f first be numerically approximated by finding the root of

$$f(a_j^f) = as_j - a_j^f - \sum_{i=1}^{m} \frac{l_i a_j^f}{a_j^f + K_i} \tag{2.18}$$

within the interval $0 < a_j^f < as_j$. This was performed using compiled C code modified from the *zeroin* C/Fortran root-finding subroutine.

Derivation of $f_{\text{cost}}(\boldsymbol{x})$

The cost function $f_{\text{cost}}(\boldsymbol{x})$ is derived from the product of the negative log multinomial probability mass function for each column *j*

$$\begin{aligned} f_{\text{cost}}(\boldsymbol{x}) &= -\ln \prod_{j=1}^{n} f_{\text{mult}}(\boldsymbol{y}_j, \boldsymbol{\pi}_j) \\ &= -\ln \prod_{j=1}^{n} \frac{Y_j! \prod_{i=1}^{m} \pi_{ij}^{y_{ij}}}{\prod_{i=1}^{m} y_{ij}!}, \end{aligned} \tag{2.19}$$

where π_{ij} is the expected frequency of each site type *i* in sample *j* according to the model values x_{ij}, and $Y_j = \sum_{i=1}^{m} y_{ij}$. Each expected frequency vector $\boldsymbol{\pi}_j$ is trivially given by $\boldsymbol{x}_j / X_j$ (where

$X_j = \sum_{i=1}^{m} x_{ij}$), thereby providing the link between the model simulation and subsequent likelihood estimation. Substituting π_{ij} and distributing the natural log yields

$$f_{\text{cost}}(\boldsymbol{x}) = \sum_{j=1}^{n}\left(Y_j \ln X_j - \sum_{i=1}^{m} y_{ij} \ln x_{ij} + \sum_{i=1}^{m} \ln y_{ij}\,! - \ln Y_j\,! \right). \tag{2.20}$$

After discarding the third and fourth terms in equation (2.20) because they do not contain any terms of $\boldsymbol{x}_j$, and are therefore not related to the MLE estimation of $\boldsymbol{\theta}$, the final cost function is given by

$$f_{\text{cost}}(\boldsymbol{x}) = \sum_{j=1}^{n}\left(Y_j \ln X_j - \sum_{i=1}^{m} y_{ij} \ln x_{ij} \right). \tag{2.21}$$

Derivation of $f_{\text{grad}}(\boldsymbol{\theta})$

The function $f_{\text{grad}}(\boldsymbol{\theta})$ returns the derivative of the cost function with respect to each component of $\boldsymbol{\theta}$:

$$\begin{aligned} f_{\text{grad}}(\boldsymbol{\theta}) &= \nabla f_{\text{cost}}(\boldsymbol{x}(\boldsymbol{\theta})) \\ &= \left(\frac{\partial f_{\text{cost}}}{\partial \theta_1},\ \frac{\partial f_{\text{cost}}}{\partial \theta_2},\ \ldots,\ \frac{\partial f_{\text{cost}}}{\partial \theta_{m+2}} \right). \end{aligned} \tag{2.22}$$

Invoking a new subscript $k \in [1\ ..\ m+2]$, we now derive an expression for each component, using the notation of $\frac{df_{\text{cost}}}{d\theta_k}$ rather than $\frac{\partial f_{\text{cost}}}{\partial \theta_k}$, reserving the $\frac{\partial}{\partial}$ notation for formalizing the isolated dependencies of x_{ij} on g_{ij}, c_{ij}, and θ_k, and of c_{ij} on a_j and θ_k, while holding all over model parameters and values constant. We derive $\frac{df_{\text{cost}}}{d\theta_k}$ using the chain rule:

$$\begin{aligned} \frac{df_{\text{cost}}}{d\theta_k} &= \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{\partial f_{\text{cost}}}{\partial x_{ij}} \frac{dx_{ij}}{d\theta_k} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{\partial f_{\text{cost}}}{\partial x_{ij}} \left(\frac{\partial x_{ij}}{\partial \theta_k} + \sum_{i'=1}^{m} \frac{\partial x_{ij}}{\partial c_{i'j}} \frac{dc_{i'j}}{d\theta_k} \right). \end{aligned} \tag{2.23}$$

$\frac{\partial f_{\text{cost}}}{\partial x_{ij}}$ is obtained by differentiating equation (2.21)

$$\frac{\partial f_{\text{cost}}}{\partial x_{ij}} = \frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}, \tag{2.24}$$

and both $\frac{\partial x_{ij}}{\partial \theta_k}$ and $\frac{\partial x_{ij}}{\partial c_{i'j}}$ are obtained by differentiation of equation (2.11)

$$\begin{aligned}\frac{\partial x_{ij}}{\partial \theta_k} &= e^{\theta_k} \frac{l_i - c_{ij}}{L - C_j} \delta_{k(m+2)} \\ &= b \frac{l - c_{ij}}{L - C_j} \delta_{k(m+2)}, \end{aligned} \tag{2.25}$$

$$\frac{\partial x_{ij}}{\partial c_{i'j}} = b \frac{l_i - c_{ij}}{(L - C_j)^2} + \left(1 - \frac{b}{L - C_j}\right) \delta_{i'i}, \tag{2.26}$$

where δ_{ab} (or equivalently $\delta_{a(b)}$) is the Kronecker delta function, defined as:

$$\delta_{ab} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \tag{2.27}$$

Substituting for $\frac{\partial f_{\text{cost}}}{\partial x_{ij}}$, $\frac{\partial x_{ij}}{\partial \theta_k}$ and $\frac{\partial x_{ij}}{\partial c_{i'j}}$ into (2.23) using (2.24), (2.25), and (2.26), respectively, and rearranging yields

$$\begin{aligned}\frac{df_{\text{cost}}}{d\theta_k} = \sum_{j=1}^{n} \frac{1}{L - C_j} \sum_{i=1}^{m} \Bigg(\left(\frac{Y_i}{X_j} - \frac{y_{ij}}{x_{ij}} \right) \times \\ \left(b(l_i - c_{ij}) \delta_{k(m+2)} + (L - C_j - b) \frac{dc_{ij}}{d\theta_k} + b \frac{l_i - c_{ij}}{L - C_j} \frac{dC_j}{d\theta_k} \right) \Bigg). \end{aligned} \tag{2.28}$$

Inspection of equation (2.28) reveals that the derivatives associated with the K_D and AGO concentrations in the reaction (i.e., $k \in [1 \,..\, m+1]$) use only the second and third terms within the last factor due to the Kronecker delta function, whereas the derivative associated with the

parameter describing the nonspecifically recovered RNA (i.e., $k = m + 2$) uses only the first term, because calculation of c_{ij} does not depend on b. Using equation (2.28) requires an expression for $\frac{dc_{ij}}{d\theta_k}$ and its sum over all site types, $\frac{dC_j}{d\theta_k}$. Application of the chain rule yields

$$\frac{dc_{ij}}{d\theta_k} = \frac{\partial c_{ij}}{\partial \theta_k} + \frac{\partial c_{ij}}{\partial a_j^f}\frac{da_j^f}{d\theta_k}, \tag{2.29}$$

and differentiation of equation (2.17) yields

$$as_j \delta_{k(m+1)} = \frac{da_j^f}{d\theta_k} + \sum_{i=1}^{m} \frac{dc_{ij}}{d\theta_k}. \tag{2.30}$$

Substituting for $\frac{da_j^f}{d\theta_k}$ in equation (2.29) with equation (2.30) results in

$$\frac{dc_{ij}}{d\theta_k} = \frac{\partial c_{ij}}{\partial \theta_k} + \frac{\partial c_{ij}}{\partial a_j^f}\left(as_j \delta_{k(m+1)} - \sum_{i=1}^{m} \frac{dc_{ij}}{d\theta_k} \right), \tag{2.31}$$

where $\sum_{i=1}^{m} \frac{dc_{ij}}{d\theta_k} = \frac{dC_j}{d\theta_k}$. This indicates that solving for $\frac{dc_{ij}}{d\theta_k}$ requires first a solution for $\frac{dC_j}{d\theta_k}$. Summing both sides of equation (2.31) for all site types $i \in [1 \,..\, m]$ yields

$$\sum_{i=1}^{m} \frac{dc_{ij}}{d\theta_k} = \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial \theta_k} + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f}\left(as_j \delta_{k(m+1)} - \frac{dC_j}{d\theta_k} \right)$$

$$\frac{dC_j}{d\theta_k} = \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial \theta_k} + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f} as_j \delta_{k(m+1)} - \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f}\frac{dC_j}{d\theta_k}, \tag{2.32}$$

Rearranging equation (2.32) yields

$$\frac{dC_j}{d\theta_k} = \frac{\sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial \theta_k} + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f} as_j \delta_{k(m+1)}}{1 + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f}} \tag{2.33}$$

For the purposes of clarity, we define

$$\phi_{ij} \equiv \frac{\partial c_{ij}}{\partial a_j^f} = \frac{l_i K_i}{(a_j^f + K_i)^2}, \tag{2.34}$$

such that

$$\frac{\partial c_{ij}}{\partial \theta_k} = \frac{-a_j^f l_i K_i}{(a_j^f + K_i)^2} \delta_{ki} = -a_j^f \phi_{ij} \delta_{ki}, \tag{2.35}$$

and we also define

$$\Phi_j \equiv \sum_{i=1}^{m} \phi_{ij}. \tag{2.36}$$

Equation (2.33) now reads as

$$\frac{dC_j}{d\theta_k} = \frac{-a_j^f \phi_{kj} \mathbb{I}_{[1..m]}(k) + \Phi_j as_j \delta_{k(m+1)}}{1 + \Phi_j}, \tag{2.37}$$

where $\mathbb{I}_{[a..b]}(x)$ is the indicator function, defined as:

$$\mathbb{I}_{[a..b]}(x) = \begin{cases} 1 & \text{if } x \in [a \,..\, b], \\ 0 & \text{if } x \notin [a \,..\, b]. \end{cases} \tag{2.38}$$

Substituting for $\frac{dC_j}{d\theta_k}$ into equation (2.31) using equation (2.37) yields

$$\begin{aligned} \frac{dc_{ij}}{d\theta_k} &= -a_j^f \phi_{ij} \delta_{ki} + \phi_{ij} \left(as_j \delta_{k(m+1)} - \frac{-a_j^f \phi_{kj} \mathbb{I}_{[1..m]}(k) + \Phi_j as_j \delta_{k(m+1)}}{1 + \Phi_j} \right) \\ &= -a_j^f \phi_{ij} \left(\delta_{ki} - \frac{\phi_{kj} \mathbb{I}_{[1..m]}(k)}{1 + \Phi_j} \right) + \frac{\phi_{ij} as_j \delta_{k(m+1)}}{1 + \Phi_j}. \end{aligned} \tag{2.39}$$

Because of complexity of equations, the full solution of $f_{grad}(\boldsymbol{\theta})$ is not shown. It is given by substituting for $\frac{dc_{ij}}{d\theta_k}$ and $\frac{dC_j}{d\theta_k}$ in equation (2.28) using equations (2.39) and (2.37), respectively. The *m*th component of the gradient is set to 0 throughout the optimization routine, which forces the value of this parameter to stay fixed at its initialized value.

Parameter initialization for relative K_D *estimation*

Each θ_i where $i \in [1, .., m]$ (i.e., $\ln(K_i)$ value) is initialized as the log of the average enrichment of that site type in each sample associated with a particular experiment:

$$\theta_{i,0} = \ln\left(\frac{1}{n} \sum_{j=1}^{n} \frac{y_{ij}}{Y_j} \Big/ \frac{y_i^l}{Y^l} \right), \tag{2.40}$$

where, as before, y_{ij} represents the read counts associated with site type *i* in sample *j*, y_i^l is the concentration of site type *i* in the RNA library, and Y_j and Y^l are their respective sums.

The initial value of the parameter θ_m is initialized and fixed at 0, which corresponds to a no-site K_D value of 1 nM. We note that fixing θ_m such that the no-site K_D value were 10 nM rather than 1 nM causes the K_D values of the other sites to also increase by 10-fold. For this reason, we report the site type K_D values as relative K_D values despite their correspondence to units of nM within the model. Finally, we initialize the parameter values of θ_{m+1} and θ_{m+2} (which correspond to the stock concentration of the AGO–miRNA complex and the concentration of nonspecific library RNA recovered in the experiment, respectively), at 2.997532 and −2.302585, corresponding to values of 20 nM and 0.01 nM, respectively. Prior to proceeding with the optimization, the values are partially randomized by adding to each parameter $\theta_{i \neq m}$ a value drawn from a normal distribution with mean 0 and standard deviation of either 0.1 or 0.01 when

optimizing K_D values for defined site lists (Figures 1–4 and S1–S6) and 12-nt *k*-mers (Figures 5, 6, and S7–S12), respectively.

Estimation of 95% confidence intervals for relative K_D values

There is no pre-existing approach for estimating the error associated with relative K_D values derived from RBNS and biochemical modeling. We devised a strategy using bootstrapping that took into account 1) error caused by sample-to-sample variation, and 2) error caused by the inherent multinomial down-sampling of RNA library molecules during sequencing. We performed the relative K_D optimization 200 times for each experiment, with each iteration i of the optimization having AGO-binding sample $j = \text{ceil}\left(\frac{i}{40}\right)$ withheld from matrix $\boldsymbol{y}$, and with the read counts in the input sequencing $\boldsymbol{y}^j$ and $\boldsymbol{y}$ resampled using the total and column-wise multinomial frequencies of each site type, respectively, with the 2.5th- and 197.5th-percentile values of each parameter used to define the plotted 95% confidence intervals. When textually reporting relative K_D values, the indicated range is given by the difference between the relative K_D value corresponding to the logarithmic mean of all 200 iterations and that of the 2.5th-percentile relative K_D value.

When calculating relative K_D values from the AGO-RBNS experiment using the first preparation of AGO2–miR-7, this procedure was modified because the stock AGO–miRNA complex was not as highly concentrated as the others, which led to decreased saturation in the higher-concentration AGO samples and therefore greater error attributable to which column j was withheld during bootstrapping. To overcome this, we first performed the optimization using all four samples, set the parameters θ_{m+1} and θ_{m+2} (corresponding to a and b) to the corresponding values estimated from this initial optimization, and fixed these values by setting their respective components of the gradient function to 0.

Read assignments

Assignment of each read to a site category was performed by searching for all possible sites within the 47-nt portion of the library molecule encompassing the 37-nt random-sequence region and 5 nucleotides of constant primer-binding sequence on either side, except in the case of miR-1. For the AGO-RBNS experiments performed with the first and second preparation of AGO2–miR-1, the libraries contained a 40-nt random-sequence region while erroneously lacking the TCG at the 5′ end of its 3′ constant sequence required for pairing to the Illumina reverse primer sequence during bridge-amplification (Data S1, Libraries 1 and 2). This caused a TCG at positions 38–40 to be near-uniformly observed in the sequencing data. We therefore restricted site identification for miR-1 to a 41-nt region corresponding to the first 36 nucleotides of the random-sequence region and the preceding five nucleotides of constant primer-binding sequence.

The procedure for estimating K_D values used only reads containing single sites. Those reads that had multiple instances of distinct sites (e.g., a read containing an 8mer site starting at position 2 of the random sequence and a 6mer site starting at position 15), as well as reads that had partially overlapping sites [e.g., a read in the miR-124 experiment containing GTGCCTT**A**<u>**AGTGTCCTT**</u>, which has an 8mer site (GTGCCTT**A**) overlapping an AA-7mer-m8bU6 site (<u>**A**</u>**AGTGTCCTT**)] were not included. When analyzing the relative affinity of all possible 11-nt registers of pairing (Figure 3A), of sites identified in Kim *et al*. (2016) (Kim et al., 2016)(Figure S3), or of sites with all possible single-nucleotide bulges and deletions (Figure S4), we identified reads that contained either an instance of the aforementioned pairing category or one of the six canonical sites, discarding any reads that contained multiple sites. Because the multisite reads made up only a small fraction (<3%) of any library, the omission of multi-site reads did not substantially distort the relative K_D values.

When calculating relative K_D values for 12-nt *k*-mers of a particular miRNA (Figures 5, 6, and S7–S12), counts from reads with more than one 12-nt *k*-mer were apportioned equally across those *k*-mers (i.e., a read containing three 12-nt *k*-mers would contribute 1/3rd to the total count of each).

Input-library sequencing

Because longer sites were rare in the input libraries, accurate quantification of their enrichment required extensive sequencing of the input libraries. To achieve the required sequencing depth, we combined sequencing results of input from experiments that used library 3. These input reads were used to assign all K_D values for let-7a, miR-155, miR-124, and lsy-6. They were also used to assign the flanking dinucleotide K_D values for miR-1.

Modeling flanking-dinucleotide effects on site K_D values

To test the consistency of the flanking-dinucleotide effect across site types and miRNAs, and to quantify the contributions of the different flanking positions, we used multiple linear regression to build a mathematical model that predicted the effect of flanking dinucleotides. The predicted affinity K_{ijk} for each combination of miRNA *i*, site-type *j*, and flanking-dinucleotide context *k* was fit as

$$K_{ijk} = \exp\left(s_{ij} + \sum_{p=1}^{4} \beta_p(n_{kp}) + \sum_{p=1}^{2} \gamma_p(d_{kp}) \right), \tag{2.41}$$

where s_{ij} is the coefficient representing the core binding affinity associated with miRNA *i* and site type *j*; $\beta_p(n_{kp})$ represents the contribution to binding of nucleotide *n* (= A, C, G, or U) at position *p* across from the four possible positions within flanking dinucleotide context *k*,

counting from the 5′ end of the target; and $\gamma_p(d_{kp})$ represents any further contribution given by the interaction of the two adjacent nucleotides making up either of the two flanking dinucleotides d (= AA, AC, …, or UU), where p = 1 or 2 refers to the 5′ and 3′ flanking dinucleotide, respectively.

Leave-one-out cross validation of this model was performed for each of the six miRNAs, leaving out the miRNA and fitting the model on the other five to obtain β_p and γ_p coefficients, using the *lm* function in R. Because the four possible nucleotide identities at each position comprised only three degrees of freedom, there was no explicit β_p coefficient for the nucleotide A, resulting in 3 × 4 β_p coefficients. For each the 5′ and 3′ flanking dinucleotides, there were correspondingly 9 γ_p coefficients describing the deviation in effect of the 9 non-A-containing dinucleotides from a linear combination of the effects of the dinucleotides that contained at least one A nucleotide, yielding a total of 9 × 2 γ_p coefficients. The plotted values and r^2 in Figure 4C (left) were calculated from the Pearson correlation coefficient describing the agreement of the observed log-transformed relative K_D values and the values predicted by the model, after normalizing all values to the average relative K_D value of the corresponding canonical site. The $\Delta\Delta G$ coefficients plotted in Figure 3 (right) are given by including a β_p of 0 for the nucleotide identity A, mean-centering the four coefficients corresponding to each position, and multiplying by RT (1.99×10^{-3} kcal K^{-1} mol^{-1} × 310.15 K).

Prediction of structural accessibility within the AGO-RBNS RNA libraries

Prediction of structural accessibility was performed by first appending each read with its appropriate 5′ and 3′ constant sequences, and folding the entire RNA library molecule in silico

using RNAplfold (Lorenz et al., 2011), with the parameters –L and –W both set to the length of the molecule, and the –u parameter set to the desired window length *w*. This produced for each read an output matrix in which the value at row *i* and column *j* corresponded to the probability that positions [*j* − *i* + 1.. *j*] are all unpaired. From this matrix the value in row *w* corresponding to a window centered on the target nucleotide pairing to miRNA position 8 or centered between those of pairing to miRNA nucleotides 7 and 8, depending on whether *w* was of odd or even length, was extracted and converted to a per-nucleotide probability by taking its *w*th root. The parameter *w* (and therefore the value after the –u flag) was either set to 15 in previous studies (Figures 4D and S6G–S6I) (Agarwal et al., 2015) or was allowed to span a range of values from 0 to 30 (Figure S6H).

RNA-seq analysis for HeLa cells

Reads were aligned to the human genome (reference assembly hg19) using STAR v2.2 with parameters –outFilterMultimapNmax 1 –outFilterMismatchNoverLmax 0.04 –outFilterIntronMotifs RemoveNoncanonicalUnannotated –outSJfilterReads Unique), and those that mapped uniquely and to ORFs were counted using htseq-count. Downstream analyses focused on the genes for which a single 3′-UTR isoform accounted for >90% of the transcripts in HeLa cells (Agarwal et al., 2015) and those with ≥10 reads in each of the libraries. The transfections were in five batches, and logTPM values were batch-normalized by fitting a linear model for each mRNA *m* to the batch identity *b* and transfected miRNA identity *t* where $\beta_{m,b}$ is the batch effect and $\beta_{m,t}$ is the batch-normalized expression value used for downstream analyses:

$$\log \mathrm{TPM}_{m,t,b} = \beta_{m,b} + \beta_{m,t}. \tag{2.42}$$

Batches were designed such that replicates for the same miRNA transfection were done in different batches.

RNA-seq analysis for HEK293FT cells

Reads were aligned as they were for RNA-seq analyses in HeLa cells. Transcript annotations were made using 3P-Seq data in HEK293 (Nam et al., 2014) to identify the genes for which a single 3′-UTR isoform accounted for >90% of the transcripts in HEK293 cells. The transfections spanned three batches, and the logTPM values were calculated and batch-normalized using equation (2.42) as per those of the HeLa transfection experiments.

Calculation of average site-type efficacy in cells

All site types identified with a relative $K_D \leq 0.1$ and represented in at least 20 instances within the 3′ UTRs of HeLa mRNAs were queried for their typical efficacy of repression in the HeLa transfection experiments (Figures 3D–3I and S6F). This was done by first calculating the repression of each mRNA *m* by miRNA *t* as

$$r_{m,t} = \beta_{m,t} - \overline{\beta_{m*,t}}, \tag{2.43}$$

where $\beta_{m,t}$ is its batch-normalized expression of in units of logTPM (see "RNA-seq analysis for HeLa cells"), and $\overline{\beta_{m*,t}}$ is its averaged expression in all other miRNA transfection experiments in which the 3′ UTR (excluding the first 15 nucleotides) contains neither an 8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-m8, or 6mer-A1 site to the guide strand nor an 8mer, 7mer-m8, 7mer-A1, or 6mer site to the passenger strand of the transfected miRNA duplex. With these $r_{m,t}$ we performed multiple linear regression

$$r_{m,t} = \sum_{j=1}^{N} n_{m,t,j} c_j, \tag{2.44}$$

where $n_{m,t,j}$ is the number of instances of site type *j* to miRNA *t* (of which there are *N* total) in the 3′ UTR of mRNA *m*, and c_j is the coefficient for the average repression conferred by site type *j*. Each coefficient c_j and corresponding 95% confidence interval were calculated using the *lm* and *confint* functions in R.

Calculation of relative K_D values for 12-nt *k*-mers

Relative K_D values for all 12-nt *k*-mers harboring at least 4 nt of complementarity to a miRNA and with the central 8 nt of the *k*-mer opposite miRNA positions 1–8 (Figures 5 and 6) were calculated as described (see "Determination of K_D values from AGO-RBNS data") over five separate batches. Each batch contained all possible 12-nt *k*-mers with a particular 4-nt complementary sequence (i.e., the first batch for miR-1 calculated the relative K_D of 12-nt *k*-mers defined by NNNNNNTCCANN, the second batch calculated that of those defined by NNNNNTTCCNNN, etc.). To minimize any systematic differences in relative K_D values calculated across the five batches, the batches were standardized by adding a constant offset (in log space) to each batch that maximized the agreement of calculated relative K_D values of *k*-mers found in more than one batch.

Biochemical model for predicting repression

Modeling AGO occupancy and mRNA repression

Given the free concentration of AGO2 loaded with miRNA g, a_g, the occupancy of the complex on a target site with a particular K_D value in the 3′ UTR of mRNA m is given by

$$\theta_{m,g,\mathrm{UTR3}} = \frac{a_g}{a_g + K_D}. \tag{2.45}$$

Because ORF sites are less efficacious than sites with the same sequence in 3′ UTRs, we fit a global penalty term c_{ORF} for sites in the mRNA ORFs:

$$\theta_{m,g,\mathrm{ORF}} = \frac{a_g}{a_g + c_{\mathrm{ORF}} K_D}. \tag{2.46}$$

Under the assumption that the binding sites act independently, an mRNA molecule with p potential binding sites for a miRNA in its ORF and q potential binding sites for a miRNA in its 3′ UTR has a miRNA occupancy of

$$N_{m,g} = \sum_{i=1}^{p} \frac{a_g}{a_g + c_{\mathrm{ORF}} K_{D,i}} + \sum_{j=1}^{q} \frac{a_g}{a_g + K_{D,j}}. \tag{2.47}$$

The background occupancy of AGO–miRNA complexes on an mRNA is estimated by substituting in the average affinity of nonspecifically bound sites (i.e., $K_D = 1.0$) for the affinity values in (2.47), ensuring that the background term is proportional to the length of the mRNA ORF and 3′ UTR.

$$N_{m,g,\mathrm{background}} = \sum_{i=1}^{p} \frac{a_g}{a_g + c_{\mathrm{ORF}}(1.0)} + \sum_{j=1}^{q} \frac{a_g}{a_g + (1.0)}. \tag{2.48}$$

For a given mRNA m and miRNA g in a transfection experiment, let $N_{m,g}$ be the occupancy of the transfected miRNA on the mRNA, α_m be the mRNA transcription rate, β_m be the portion of

the mRNA decay rate that is not due to the transfected miRNA, and *b* represent the amplification of the decay rate introduced by the binding of one AGO–miRNA complex. We model the abundance of the mRNA in transfected cells, $y_{m,g}$, according to its transcription rate and aggregate decay rate:

$$\frac{dy_{m,g}}{dt} = \alpha_m - \beta_m(1 + bN_{m,g})y_{m,g}. \tag{2.49}$$

At steady-state, the abundance of the mRNA in transfected cells is therefore

$$y_{m,g} = \frac{\alpha_m}{\beta_m(1 + bN_{m,g})}. \tag{2.50}$$

In the absence of the transfected miRNA, the steady-state abundance of the mRNA would be

$$y_{m,0} = \frac{\alpha_m}{\beta_m(1 + bN_{m,g,\text{background}})}. \tag{2.51}$$

The fold-change *r* caused by the transfected miRNA is therefore

$$r_{m,g} = \frac{y_{m,g}}{y_{m,0}} = \frac{1 + bN_{m,g,\text{background}}}{1 + bN_{m,g}}. \tag{2.52}$$

We assumed that TPM values for a given transcript follow a log-normal distribution, so the fitting was done using log(expression) and log(fold change) values:

$$\log r_{m,g} = \log(1 + bN_{m,g,\text{background}}) - \log(1 + bN_{m,g}). \tag{2.53}$$

Fitting the biochemical model to RNA-seq measurements

To measure $y_{m,0}$, and thus $r_{m,t}$, it is common to measure mRNA abundances after performing a mock transfection. However, mock transfections often introduce their own systematic gene

expression changes and fail to capture the derepression signal from endogenous miRNAs that is observed upon miRNA transfection (Saito and Sætrom, 2012). To avoid these complications, we took advantage of the observation that we do not explicitly need this value to fit the model with the assumption that $y_{m,0}$ does not change between different transfection experiments (i.e., the basal decay rates of mRNAs not bound by transfected miRNAs are unchanged between transfection experiments). Under this assumption, we can fit mean-centered expression values against mean-centered repression values. Consider the repression of mRNA *m* by miRNA *g* out of *G* miRNA transfection experiments,

$$\begin{aligned}
\log r_{m,t} - \overline{\log \mathbf{r}_m} &= (\log y_{m,t} - \log y_{m,0}) - \frac{1}{T}\sum_{i=1}^{T}(\log y_{m,i} - \log y_{m,0}) \\
&= (\log y_{m,t} - \log y_{m,0}) - \frac{1}{T}\sum_{i=1}^{T}\log y_{m,i} + \frac{1}{T}\sum_{i=1}^{T}\log y_{m,0} \\
&= (\log y_{m,t} - \log y_{m,0}) - \frac{1}{T}\sum_{i=1}^{T}\log y_{m,i} + \log y_{m,0} \\
&= \log y_{m,t} - \frac{1}{T}\sum_{i=1}^{T}\log y_{m,i} \\
&= \log y_{m,t} - \overline{\log \boldsymbol{y}_m}.
\end{aligned} \tag{2.54}$$

For *M* mRNAs and *G* miRNAs, we minimized the following loss function with respect to the parameters *b*, a_g, and c_{ORF}, where $\hat{\mathbf{r}}$ are the predicted repression values and **y** are the measured expression values:

$$L = \sum_{m=1}^{M}\sum_{g=1}^{G}\left((\log y_{m,g} - \overline{\log \mathbf{y}_m}) - (\log \hat{r}_{m,g} - \overline{\log \hat{\mathbf{r}}_m})\right)^2. \tag{2.55}$$

These values were used to calculate the r^2 values. For plotting fold-change values, we extrapolated the values for $y_{m,0}$ by finding the intercept of the linear relationship between the

predicted repression values and the measured expression values (Figures 5C–5E) for each mRNA. To prevent extreme intercepts in the limit of no variability in the predicted repression, a weak Bayesian prior of $\mathcal{N}(0,\ 0.01\times\sigma^2)$ was applied to the slope estimate, where σ^2 is the variance of the error of the linear fit. This causes a transcript with very little predicted miRNA binding to any of the transfected miRNAs to have baseline values that approach the average expression of the transcript in all the transfection experiments.

Calculating features for the biochemical+ model

For each 12-nt *k*-mer in an mRNA, its raw structural-accessibility score was calculated using RNAplfold (Lorenz et al., 2011) with the flags –L 40 –W 80 –u 15 and taking the $\log_{10}$ value of the unpaired probability for a 14-nt region centered on the match to miRNA nucleotides 7 and 8 (Agarwal et al., 2015). Because the K_D values already reflect the average structural accessibility of a 12-nt *k*-mer in random contexts, the raw RNAplfold output for each site in its endogenous context was then offset by the average RNAplfold output of the same site in 200 random 40-nt contexts. Folding 200 random contexts for all 12-nt *k*-mers was laborious, so this process was only carried out for the 12-nt *k*-mers containing one of the six canonical sites. For all other 12-nt *k*-mers, the average structural accessibility for canonical sites to the same miRNA was used.

For each 12-nt *k*-mer in an mRNA containing a canonical site, the 3′-supplementary pairing score was calculated as previously (Grimson et al., 2007). This score was set to 0.0 for 12-nt *k*-mers without a canonical site. PCT values were calculated for each 12-nt *k*-mer in an mRNA 3′ UTR containing a 7mer-m8, 7mer-A1, or 8mer site using multiple alignments from 84 species as previously (Agarwal et al., 2015). This score was set to 0.0 for all other sites.

Calculating site occupancy in the biochemical+ model

All the additional features modified the K_D values linearly in log space (e.g., linear in ΔG space). For each 12-nt *k*-mer with $K_{D,i}$, structural-accessibility score SA_i, 3′ supplementary pairing score Threep_i, and PCT score PCT_i,

$$\log K_{D,i,\mathrm{biochem+}} = \log K_{D,i} + c_{\mathrm{SA}}\mathrm{SA}_i + c_{\mathrm{Threep}}\mathrm{Threep}_i + c_{\mathrm{PCT}}\mathrm{PCT}_i, \quad (2.56)$$

where c_{SA}, c_{Threep}, and c_{PCT} were fit alongside the other parameters (a_g, b, and c_{ORF}) fit in the biochemical model.

Refitting TargetScan7

The original TargetScan7 model (Agarwal et al., 2015) was only trained on miRNA–mRNA pairs where the miRNA had a single 6mer, 7mer-A1, 7mer-m8, or 8mer site to the mRNA 3′ UTR. This may have biased the training set towards mRNAs with short 3′ UTRs. When predicting scores for mRNAs with multiple sites, scores for the individual sites were summed. To allow TargetScan7 to be trained on all mRNAs, we fit the loss function given in (2.55) using the 16 transfection experiments of miRNA duplexes into HeLa cells.

Combined CNN and biochemical model

CNN architecture

The CNN architecture was as described in Figure S9A, with two convolutional layers and two fully connected layers. The first fully connected layer could, in principle, take into account every register of interaction between the miRNA and target sequences, including large bulges in either sequence that would significantly offset the register of pairing. However, we did not expect these types of sites to have higher-than-background binding affinities, so we applied a mask to this layer such that all interactions that would require more than a 4-nt offset in register were not

considered. This improved convergence time without affecting predictive performance during cross-validation.

Input data and training

The training dataset contained RBNS data for six miRNAs, repression data for five of those miRNAs, and repression data for 11 additional miRNAs. Because the relative K_D values for all the 12-nt *k*-mers were heavily skewed towards low-affinity sites, we increased the probability of sampling a high-affinity site during training. To do this, we assigned the 12-nt *k*-mers to bins by rounding their log K_D values to the nearest 0.25. We then assigned a weight to all the 12-nt *k*-mers in a bin such that their weighted sum would not exceed 2000 (i.e. 12-nt *k*-mers in highly populated bins received lower weights). During training, 12-nt *k*-mers were sampled according to their weights. We initially trained the model 11 times, each time leaving out one of the 11 additional transfection datasets, training on the six RBNS datasets and the 15 remaining transfection datasets, and testing on the held-out datasets. This 11-fold cross-validation allowed us to pick optimal hyperparameters. The final model was then trained on all six RBNS datasets and all 16 transfection datasets. Each mini-batch consisted of 1) RBNS measurements for 50 pairs of miRNAs and 12-nt *k*-mers and 2) repression data for 16 mRNAs for all 16 miRNAs. The ten RBNS inputs were passed through the CNN to produce predicted logK_D values, which were then compared to the measured logK_D values for those RBNS inputs to calculate the RBNS loss:

$$L_{\text{rbns}} = \sum_{i=1}^{10} (\log K_{D,i} - \log \hat{K}_{D,i})^2. \tag{2.57}$$

For each of the 32 miRNAs (two miRNA sequences for each of the 16 transfected duplexes), all 12-nt *k*-mers with at least four contiguous nucleotides of the 8mer site to the 16

miRNAs were extracted from their 3′-UTR and ORF sequences. For 12-nt *k*-mers for the same miRNA that overlapped, the 12-nt *k*-mer with the higher priority match to the 8mer site was chosen. The priority order for the match was match2–5 > match3–6 > match1–4 > match4–7 > match5–8. All of the miRNAs and 12-nt *k*-mers were passed through the same CNN as above to produce predicted K_D values. These K_D were then combined for all 12-nt matches to guide and passenger strand sequences of a transfected duplex on the same mRNA according to the biochemical model to produce predicted log fold-change values. These predictions were used to calculate the repression loss term, as in equation (2.55). Here, *g* enumerates the 16 miRNAs in the training set, *m* enumerates the 16 mRNAs in the mini-batch, and $n_{m,g}^{\text{guide,ORF}}$, $n_{m,g}^{\text{pass,ORF}}$, $n_{m,g}^{\text{guide,3′UTR}}$, and $n_{m,g}^{pass,\text{3′UTR}}$ represents the number of 12-nt matches in the ORF or 3′ UTR of mRNA *m* to the guide or passenger strands, respectively, of miRNA *g*:

$$\theta_{m,g}^{\text{guide,ORF}} = \sum_{i=1}^{n_{m,g}^{\text{guide,ORF}}} \frac{\hat{a}_g^{\text{guide}}}{\hat{a}_g^{\text{guide}} + c_{\text{ORF}} \hat{K}_{D,i}}, \quad \theta_{m,g}^{\text{guide,UTR3}} = \sum_{i=1}^{n_{m,g}^{\text{guide,UTR3}}} \frac{\hat{a}_g^{\text{guide}}}{\hat{a}_g^{\text{guide}} + \hat{K}_{D,i}}$$

$$\theta_{m,g}^{\text{pass,ORF}} = \sum_{i=1}^{n_{m,g}^{\text{pass,ORF}}} \frac{\hat{a}_g^{\text{pass}}}{\hat{a}_g^{\text{pass}} + c_{\text{ORF}} \hat{K}_{D,i}}, \quad \theta_{m,g}^{\text{pass,UTR3}} = \sum_{i=1}^{n_{m,g}^{\text{pass,UTR3}}} \frac{\hat{a}_g^{\text{pass}}}{\hat{a}_g^{\text{pass}} + \hat{K}_{D,i}}$$

$$r_{m,g} = \frac{1}{1 + b(\theta_{m,g}^{\text{guide,ORF}} + \theta_{m,g}^{\text{guide,UTR3}} + \theta_{m,g}^{\text{pass,ORF}} + \theta_{m,g}^{\text{pass,ORF}})} \tag{2.58}$$

$$L_{\text{repression}} = \sum_{m=1}^{16} \sum_{g=1}^{16} ((\log y_{m,g} - \overline{\log \mathbf{y}_m}) - (\log \hat{r}_{m,g} - \overline{\log \hat{\mathbf{r}}_m}))^2. \tag{2.59}$$

The total loss was calculated as a weighted sum of the two loss terms, along with an L_2 regularization term on the CNN weights ($\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$). Because the transfected miRNAs are expected to have similar a_g values, an L_2 regularization term was also applied to the differences between guide-strand a_g values and the average guide-strand a_g value to prevent these values from drifting too far apart initially.

$$\mathbf{d}^{\text{guide}} = \mathbf{a}_t^{\text{guide}} - \overline{\mathbf{a}^{\text{guide}}} \tag{2.60}$$

The RBNS loss weight, repression loss weight, CNN weight regularizer, and the a_g offset weights are λ_k, λ_r, λ_w, and λ_d respectively.

$$L_{\text{total}} = \lambda_k L_{\text{rbns}} + \lambda_r L_{\text{repression}} + \lambda_w(\left\|\mathbf{w}_1\right\|_2 + \left\|\mathbf{w}_2\right\|_2 + \left\|\mathbf{w}_3\right\|_2 + \left\|\mathbf{w}_4\right\|_2) + \lambda_d \left\|\mathbf{d}^{\text{guide}}\right\|_2). \tag{2.61}$$

The model was implemented in TensorFlow and trained by minimizing the total loss using the Adam optimizer with an initial learning rate of 0.003 for 100 epochs and $\lambda_k = 0.05, \lambda_r = 0.95, \lambda_w = 0.0001, \lambda_d = 0.001.$ The CNN weights were initialized randomly using Xavier initialization.

Evaluation of CNN predictions on the test set of miRNAs transfected into HEK293FT cells

For each miRNA in the test set, we generated the complete list of 262,144 12-nt *k*-mers with at least 4 nt of complementarity to the miRNA and predicted their K_D values using the CNN. To identify high-affinity noncanonical sites, we isolated the 12-nt *k*-mers without canonical sites to the miRNA, grouped them based on the 8-nt sequences centered in each 12-nt sequence, and sorted each group. Out of 64 possible 12-nt *k*-mers sharing the same 8-nt center sequence, if the 32 *k*-mers with the highest predicted affinity values contained the same 9-nt sequence encompassing the 8-nt centered sequence, the 9-nt sequence was identified as a site and assigned the average K_D value of 12-nt *k*-mers with that 9-nt sequence. Otherwise, the 8-nt sequence was identified as a site and assigned the average K_D value of 12-nt *k*-mers with that 8-nt sequence. In either case, the 12-nt *k*-mers with the new site were removed from the pool, and the processed repeated. Afterwards, only new sites with an average predicted $\ln(K_D) < -2$ (equivalent to $\log_{10}(K_D) < -0.87$) were kept. These sites were further consolidated into shorter 7-nt sequences if

several versions of the 7-nt sequence appeared in the new site list with a different flanking nucleotide. The average site-type efficacy in cells for all the canonical and annotated noncanonical sites for each miRNA was calculated as in the section "Calculation of average site-type efficacy in cells."

Predictions of miRNA–target interaction energy using other methods

To calculate the free-energy of binding for canonical site types to each miRNA (Figure 6C), the RNAduplex program (Lorenz et al., 2011) was supplied the site sequence and miRNA sequence. The predicted free-energies were reported in units of kcal/mol. To calculate MIRZA scores, we downloaded the MIRZA (Khorshid et al., 2013) algorithm from http://www.clipz.unibas.ch/mirzag/. The algorithm was run with the option to update priors and was supplied each miRNA sequence and 1000 examples of each canonical site in random 40-nt contexts (sequences of equal length between 30 and 55 nt were required). The algorithm also required relative miRNA abundances, but because each miRNA was evaluated separately, this was set to 1000 arbitrarily and did not affect output. The reported scores were the average score for the 1000 examples of each site type.

Processing of and model evaluation on external datasets

mRNA fold change data for let-7c transfection into HCT116 cells (Linsley et al., 2007), miR-124 and miR-7 transfections into HEK293 cells (Hausser et al., 2009), and miR-302/367 knockdown in hESC cells (Lipchina et al., 2011) were obtained as described (Agarwal et al., 2015). For gene-expression changes upon knockout of miR-122 in mouse liver cells (Eichhorn et al., 2014), raw RNA-Seq reads were downloaded from the GEO (GSE61073), aligned to the mouse genome mm10, and annotated using the set of representative transcripts curated in TargetScanMouse v7.1

(Agarwal et al., 2015) (http://www.targetscan.org/mmu_71). We required mRNA expression levels to exceed 10 TPM in either the wildtype or knockout samples.

Top targets identified by crosslinking experiments upon transfection of miR-124 or miR-7 into HEK293 cells (Hafner et al., 2010), knockout of miR-155 in mouse T cells (Loeb et al., 2012), and knockdown of miR-302/367 in hESC cells (Lipchina et al., 2011) were obtained as in (Agarwal et al., 2015). Gene expression changes and eCLIP-identified targets upon overexpression of miR-20a in HeLa cells (Zhang et al., 2018) were kindly provided to us by the authors.

For each dataset, biochemical and biochemical+ model predictions were generated by using global biochemical parameters fit using the transfection data into HeLa cells. For the let-7c, miR-124, miR-7, and miR-155 datasets, experimentally-determined relative K_D values (see "Calculation of relative K_D values for 12-nt *k*-mers") were used, whereas CNN-predicted K_D values were used for the miR-302/367, miR-122, and miR-20a datasets. When predicting mRNA changes upon miR-155 knockout in mouse T cells, the average a_g value of passenger strands fit for the HeLa transfection datasets was used. For all other datasets, the average a_g value of miRNA strands fit for the HeLa transfection datasets was used.

Estimation of maximal r^2 values

For each transfection experiment, we define the following random variables:

X : Direct log fold-change values, must be negative, distribution unknown
$E_1 \sim \mathcal{N}(0,\ \sigma_1^2)$: Reproducible symmetrical variability (e.g. secondary effects)
$E_2 \sim \mathcal{N}(0,\ \sigma_2^2)$: Technical/experimental noise
$Y = X + E_1 + E_2$: Observed repression values

The goal is to determine the variance of *X* compared to the variance of *Y*. While the distribution of *X* is unknown, we can approximate it using a discrete distribution with *m* discrete bins spanning the range of realistic log repression values $\boldsymbol{w} = [w_1, w_2, \ldots, w_m]$ with probabilities $\boldsymbol{p} = [p_1, p_2, \ldots, p_m]$. In practice, we used 50 bins spanning –3 to 0 in log space (–4.33 to 0 in $\log_2$ space). To calculate the probability of observing the measured repression values $\boldsymbol{y}_{1,2\ldots n} \sim Y$ given ($\sigma_1^2 + \sigma_2^2$), $\boldsymbol{w}$, and $\boldsymbol{p}$

$$\begin{aligned} \log p(\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{w}, \sigma_1, \sigma_2) &= \log \prod_{i=1}^{n} p(y_i \mid \boldsymbol{p}, \boldsymbol{w}, \sigma_1, \sigma_2) \\ &= \sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{p}, \boldsymbol{w}, \sigma_1, \sigma_2) \\ &= \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} p_j \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-(y_i - w_j)^2 / 2(\sigma_1^2+\sigma_2^2)} \right). \end{aligned} \tag{2.62}$$

We then fit values for ($\sigma_1^2 + \sigma_2^2$) and $\boldsymbol{p}$ by maximizing the likelihood of observing the data $\boldsymbol{y}$ using tensorflow.contrib.opt.ScipyOptimizerInterface(method="SLSQP") under the constraint that $\sum p_j = 1$. We estimated σ_2^2, and thus σ_1^2, by examining the reproducibility between two biological replicates

$$\sigma_2^2 \approx Var(Y_{rep1} - Y_{rep2}) / 2 \tag{2.63}$$

$$\sigma_1^2 = (\sigma_1^2 + \sigma_2^2) - \sigma_2^2 \tag{2.64}$$

and estimated the expected value and variance of *X* given $\boldsymbol{w}$ and $\boldsymbol{p}$:

$$E(X) \approx \frac{1}{m} \sum_{j=1}^{m} p_j w_j \tag{2.65}$$

$$Var(X) \approx \frac{1}{m} \sum_{j=1}^{m} p_j (w_j - E(X))^2 \tag{2.66}$$

The estimated maximal r^2 value is given by dividing $Var(X) / Var(Y)$.

Supplementary figures and tables

A

miRNA	AGO2–miR-1	AGO2–miR-155
miR-1	995,485	7,600
miR-155	164	977,308
miR-221	62	5,311
miR-17	736	563
miR-7–1	479	98
miR-320a–d	176	959
let-7a–c	270	189
miR-20a	252	171
miR-222	50	901
miR-18a–b	144	182
miR-155-3p	0	883
miR-93	135	87
Other miRNAs	2,048	5,747

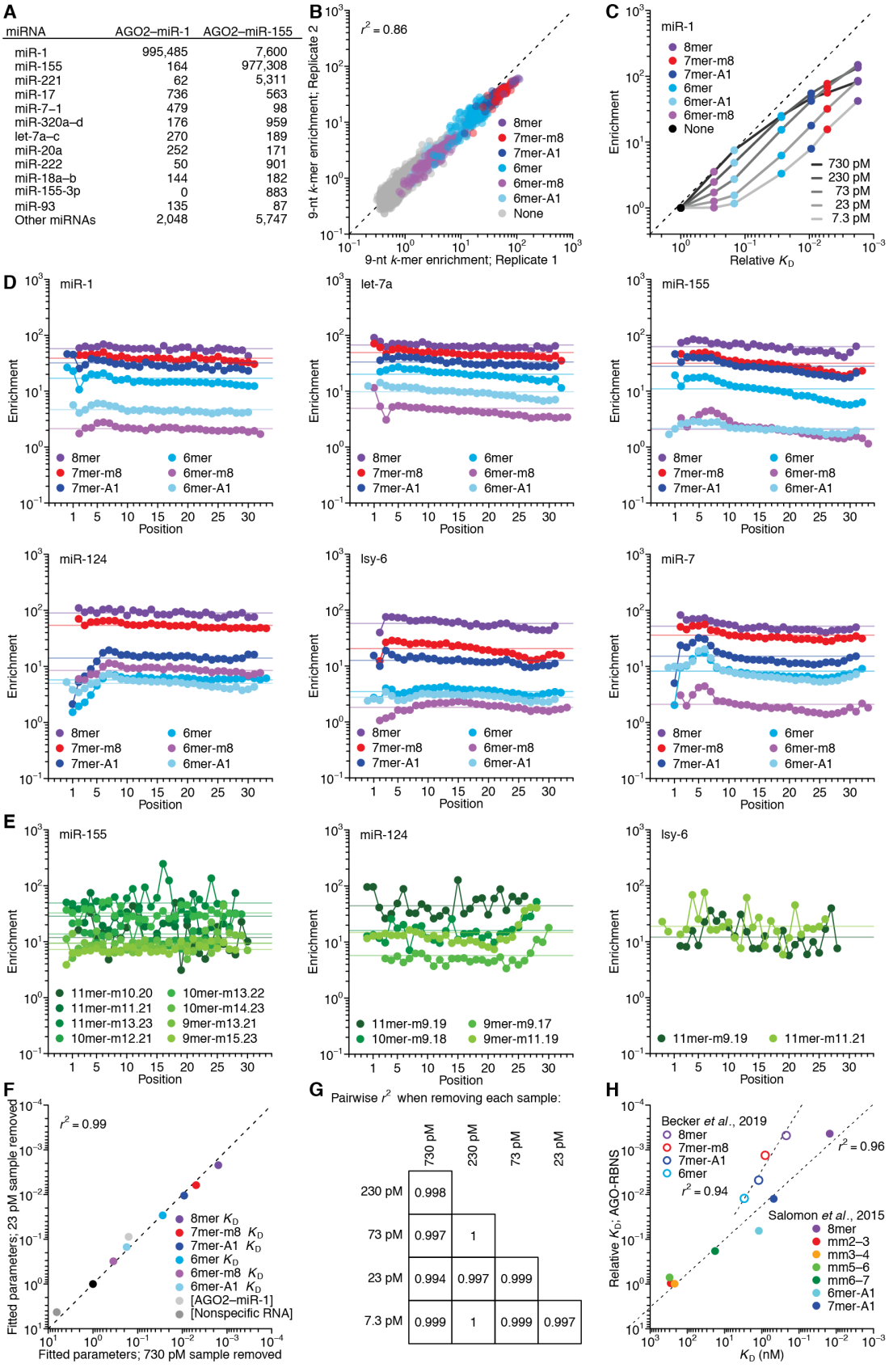


Figure S1. Reproducibility of AGO-RBNS results. (**A**) MicroRNAs observed in AGO2–miR-1 and AGO2–miR-155 preparations, as quantified using small-RNA sequencing. Shown are the counts per million mapped miRNA reads for miR-1, miR-155, and contaminating miRNAs, listing the ten most abundant contaminants observed when averaging the counts of the two samples. (**B**) Correspondence between the results of two independent AGO-RBNS binding reactions that used different preparations of purified AGO2–miR-1 and different RNA libraries, with each library generated from a different DNA synthesis. Compared is the enrichment of all 9-nt *k*-mers that contain either 8mer (purple), 7mer-m8 (red), 7mer-A1 (blue), 6mer (cyan), 6mer-m8 (violet), or 6mer-A1 (light blue) sites, as well as the enrichment of 10,000 arbitrarily chosen 9-nt *k*-mers not containing any of these sites (gray). The r^2 was calculated using the log-transformed values. The dashed line shows $y = x$. (**C**) Relationship between affinity and AGO-RBNS enrichment. The enrichments of reads containing each of the six canonical sites in addition to no-site reads (Figure 1D) are plotted their corresponding relative K_D values, for each of the five AGO2–miR-1 concentration samples. Grayscale lines denote each sample, with the 7.3 pM and 730 AGO2–miR-1 samples in light gray and black, respectively. Enrichments are normalized to that of the no-site reads in each sample. (**D**) Enrichment of canonical sites as at each position within the library molecules. Random-sequence positions are numbered from the 5′ end with respect to the 30 possible positions of an 8mer site. Points represent enrichment of the indicated canonical site (key) at each position for the most-concentrated AGO2–miRNA sample within each AGO-RBNS experiment. The high enrichments persisting in the 5′-most positions of the random-sequence region, where the miRNA 3′ region is opposite the non-complementary primer-binding sequence and therefore cannot paired, suggested minimal influence of 3′-supplementary pairing on the enrichments further 3′. Also, while neighboring primer-binding sequence sometimes had a modest influence at one end of the random-sequence region, this had a negligible effect on the overall enrichment observed for each site type (horizontal lines). (**E**) Enrichment of 3′-only sites as a function of their position within the library molecules. Random-sequence positions are numbered with respect to the 27 possible positions of an 11-nt site. Otherwise, as in (D). When analyzing the uniformity of enrichment of canonical (D) and 3′-only sites (E), we identified reads that contained only a single instance of a site, considering all the sites identified by *k*-mer enrichment analysis (supplemented with the 6mer-m8 site in the case of miR-7), all single-nucleotide mismatch variants of the 8mer, the 7mer-m8, the 7mer-A1, and the 6mer, and the four contiguous 5mer sites within the seed region (i.e., the 5mer-A1, 5mer-m2.6, 5mer-m3.7, and the 5mer-m8 sites). This was to ensure that the positional site enrichments detected were not influenced by the presence of any weaker sites elsewhere within the read. (**F** and **G**) Robust estimation of relative K_D values and other parameters. To estimate the uncertainty of the fitted model parameters (key), the MLE procedure was repeated five times, each time excluding data from one of the five AGO2–miR-1 concentrations. The Pearson r^2 was calculated between each of the 10 pairwise possibilities as in (F), which shows the comparison of the least well correlated pair (that when omitting the 23 and 730 pM AGO2–miR-1 samples, respectively) (dashed line, $y = x$). All ten pairwise comparisons are reported in (G). (**H**) The correspondence between the relative K_D values determined by AGO-RBNS with K_D values reported by two prior studies (Becker et al., 2019; Salomon et al., 2015). Plotted are values for the indicated sites to let-7a (key). To account for the potential effects of flanking nucleotides in the target RNAs of Salomon *et al*. (Salomon et al., 2015), for each comparison we use the relative K_D value of the 12-nt *k*-mer that contains the site and flanking sequence context of the corresponding target RNA. Because each of the four canonical-site K_D values reported in Becker *et al*. (2019) (Becker

et al., 2019)represents the median for multiple target RNAs containing that site, for each comparison we use the relative K_D value of the site determined without consideration of flanking sequences (Figure 2A).

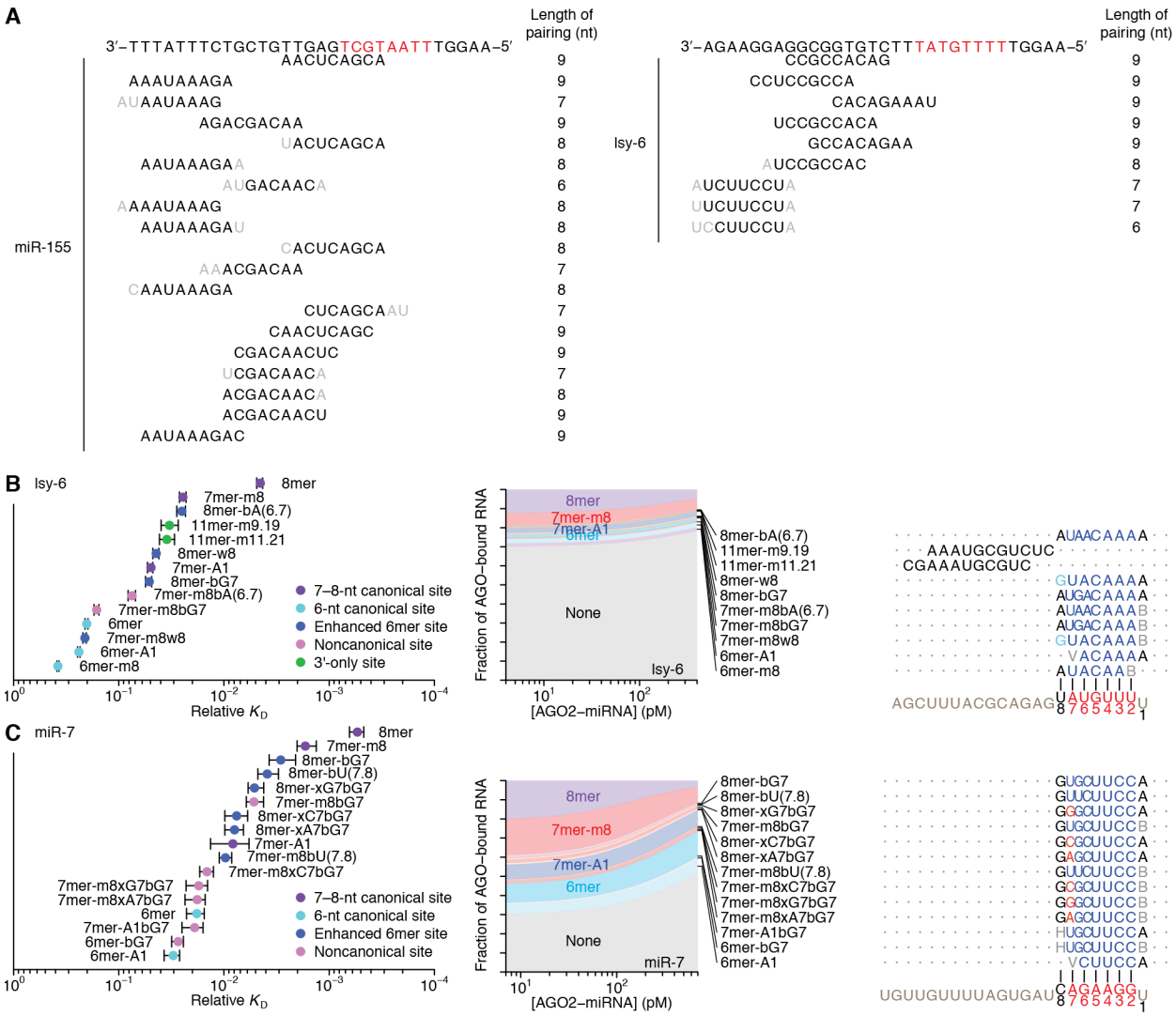


Figure S2. Additional sites identified through AGO-RBNS. (**A**) Enriched motifs that were identified for miR-155 and lsy-6 yet lacked complementarity to the respective guide sequence, aligned to highlight their complementarity to the competitor oligo used to purify the AGO–miRNA complex. Because these motifs each had ≥6 nt of complementarity to the competitor oligo and relatively little complementary to the miRNA, they were excluded as sites to the miRNA. The red nucleotides indicate the region of the competitor oligo that is identical to positions 1–8 of the miRNA. (**B** and **C**) Relative K_D values and proportional occupancy of established and newly identified sites of lsy-6 (B) and miR-7 (C), as in Figure 2. The identified sites, their relative K_D values with 95% confidence intervals, and the enriched 10-nt *k*-mers used for iterative site identification, are reported in Data S2. These analyses also detected an AACGAGGA motif for lsy-6 and a GCUUCCGC motif for miR-7, which were assigned relative K_D values of $1.58 \pm 0.07 \times 10^{-1}$ and $1.1 \pm 0.5 \times 10^{-2}$, respectively. These two motifs were not considered miRNA sites because each did not match its respective miRNA and each did not mediate repression in our reporter assays (Figure S5B).

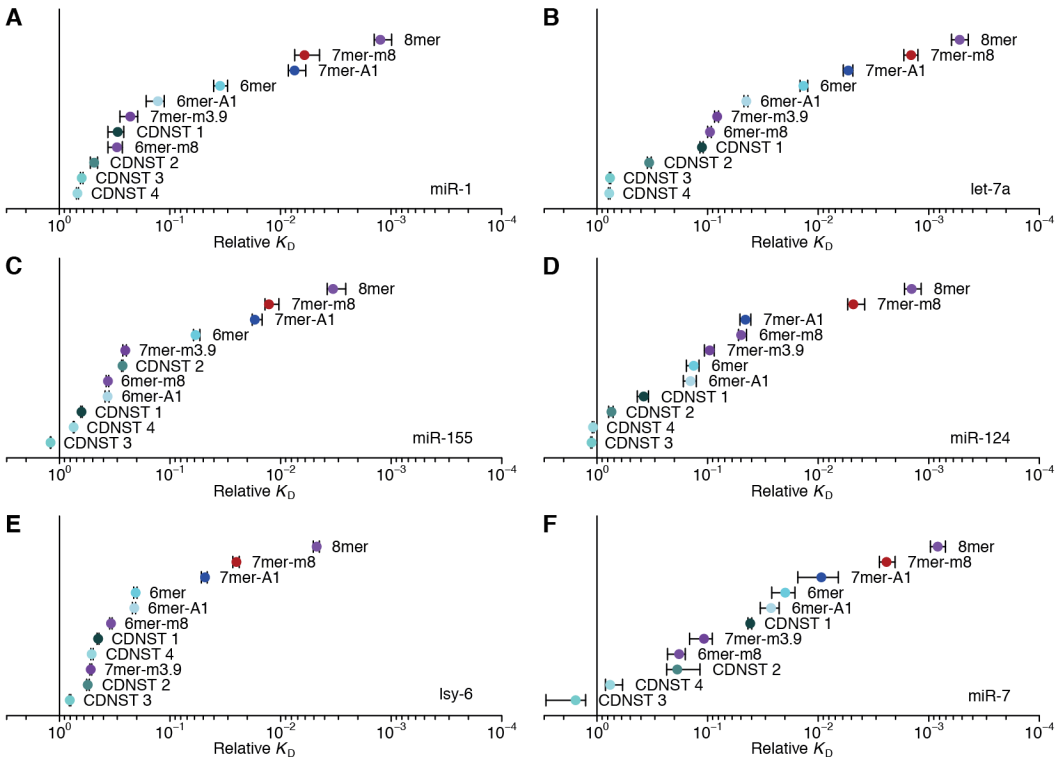


Figure S3. Relative K_D values of site types reported in Kim *et al.* (2016). (**A**–**E**) Analysis was as in Figure 1F but performed using the site types of Kim et al., (2016)(Kim et al., 2016), which include the canonical sites (Figure 1A), an offset 7mer (which pairs to miRNA nucleotides 3–9), as well as four context-dependent noncanonical site types (CDNST) that are proposed to substantially extend the scope of miRNA–mRNA regulatory interactions. The offset 7mer site bound with similar affinity as its nested 6mer-m8 site, with effects of flanking nucleotide composition (Figure 4) explaining any minor differences. The context-dependent noncanonical site type 1 (CDNST 1) pairs to miRNA nucleotides 2–6 and lacks both a match at position 7 and an A at target position 1 (equivalent to the 5mer-m2.6 site); for each miRNA, this site bound better than no site, and for miR-1, and let-7a its affinity exceeded the thresholds for site identification in our analyses, conferring 3.6- and 9.5-fold greater affinity over no site–containing reads, respectively (Figures 1F and 2A). This site was also detected in analysis of our first miR-7 replicate (Data S2). CDNST 2 is a 7mer-A1 site with a mismatch at position 5; this site includes the 7mer-A1xU5 site identified for miR-155 (Figure 2B), but otherwise bound with affinity below the thresholds of our analyses. CDNST 3 and CDNST 4, which each have three mismatches to the seed, bound with affinity resembling that of no site. For each CDNST with an internal mismatch, the relative K_D value represents the aggregate value for all mismatched variants.

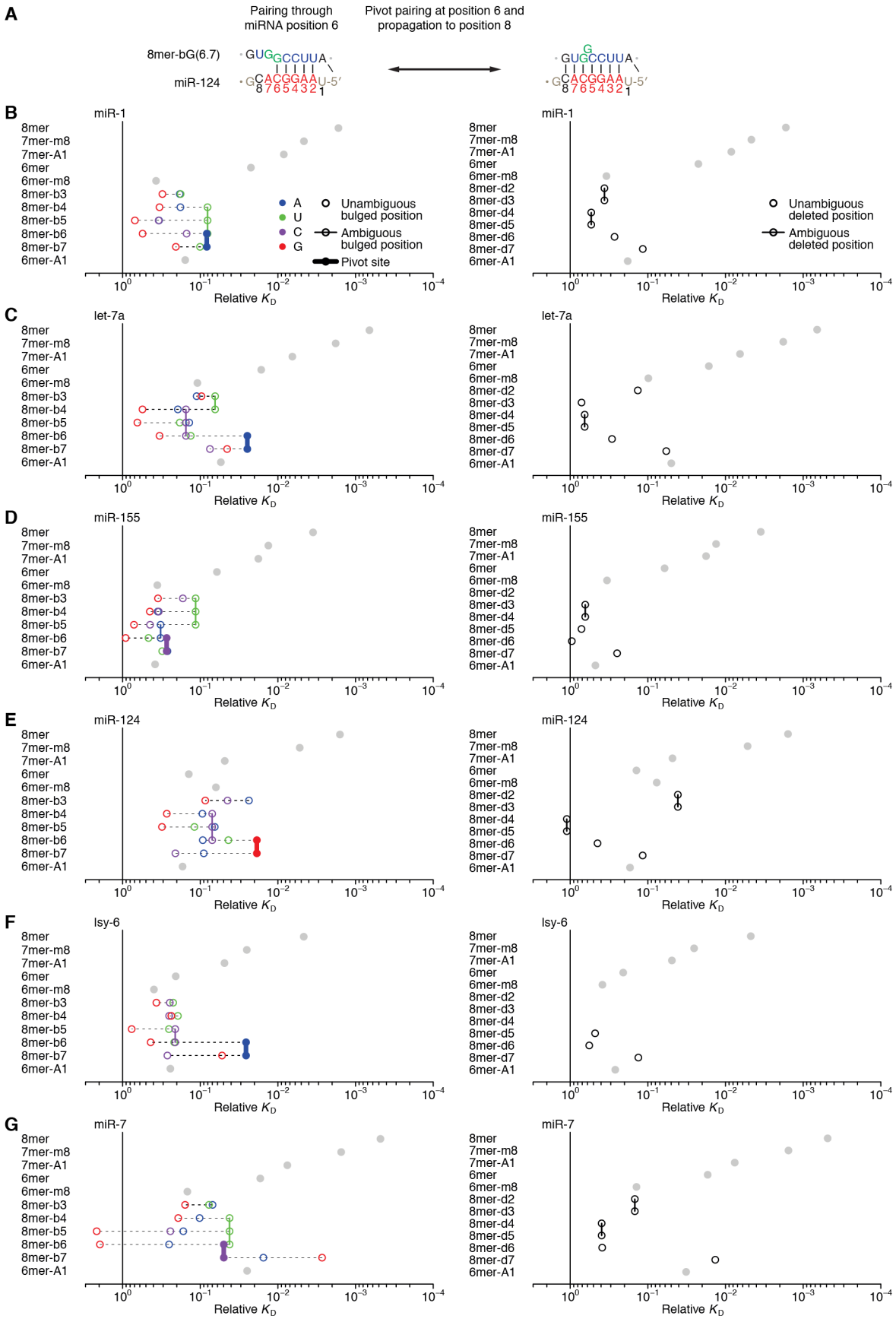

A
Pairing through miRNA position 6
Pivot pairing at position 6 and propagation to position 8
8mer-bG(6.7)
miR-124
B
miR-1
C
let-7a
D
miR-155
E
miR-124
F
lsy-6
G
miR-7
8mer
7mer-m8
7mer-A1
6mer
6mer-m8
8mer-b3
8mer-b4
8mer-b5
8mer-b6
8mer-b7
6mer-A1
8mer-d2
8mer-d3
8mer-d4
8mer-d5
8mer-d6
8mer-d7
A
U
C
G
Unambiguous bulged position
Ambiguous bulged position
Pivot site
Unambiguous deleted position
Ambiguous deleted position
Relative K_D

Figure S4. Analysis of the effects of bulged nucleotides. (**A**) The proposed pathway for paring between miR-124 and its pivot site (or 8mer-bG(6.7)) (Chi et al., 2012). For pivot sites, the target nucleotide that pairs to miRNA nucleotide 6 is repeated to create a bulge that ambiguously maps to positions 6 or 7. (**B**–**G**) Relative K_D values examining the effect of a bulged target nucleotide (left) or a bulged miRNA nucleotide (right) within a site to either miR-1 (B), let-7a (C), miR-155 (D), miR-124 (E), lsy-6 (F), or miR-7 (G). Analysis was as in Figure 1F but values are plotted for 8mer sites with a bulged or deleted nucleotide (left and right, respectively), as indicated in each key. Values for the six canonical sites are also plotted for reference (filled gray circles). Dashed horizontal lines connect points for different bulged nucleotides at the same position. Points representing bulged or deleted nucleotides at ambiguous positions are connected with vertical lines. For example, three green points showing the result for ACAUUUCCA (a miR-1 site that has a bulged U at either target positions 4, 5, or 6) are connected with a green line in (A). Some of the sites with ambiguous bulged positions are classified as pivot sites (Chi et al., 2012), (e.g., the ACAAUUCCA site for miR-1); points representing pivot sites are filled and connected with a wide vertical lines. Although the pivot sites for miR-124 and lsy-6 bound with affinities substantially exceeding those of their nested 6mer-A1 sites and were thus identified as unique sites in our analysis [Figure 2, 8mer-bG(6.7) and 8mer-bA(6.7), respectively], pivot sites for the other miRNAs bound with affinities resembling those of their nested 6mer-A1 sites, with effects of flanking nucleotide composition (Figure 4) explaining any minor differences [e.g., the let-7a 8mer-bA(6.7) sequence CUAACCUCA also corresponds to a 6mer-A1 (underlined) with a favorable UA dinucleotide context]. Moreover, for miR-1 [8mer-bU(4.6)], miR-155 [8mer-bU(3.5)] and miR-7 (8mer-bG7), other types of bulged sites bound substantially better than did the pivot sites.

The pivot site is proposed to mediate widespread targeting (Chi et al., 2012). This noncanonical site has canonical pairing to the seed region, except that the target residue matching position 6 of the miRNA is repeated, which forces a single-nucleotide bulge at position 6 or 7 of the target (Chi et al., 2012). Our de novo search for sites supported pivot sites of miR-124 and lsy-6. For example, the miR-124 8mer-bG(6.7) site (an 8mer site but with an extra G bulged at either position 6 or 7) is a 9-nt pivot site with affinity exceeding that of the canonical 7mer-A1 site, and the lsy-6 8mer-bA(6.7) is a 9-nt pivot site with affinity matching that of the canonical 7mer-m8 site (Figures 2C and S2B). However, even though these pivot sites for miR-124 and lsy-6 were among the highest-affinity noncanonical sites identified, we did not identify pivot sites for any of the other four miRNAs (Figures 1F, 2A, 2B, and S2C), and a systematic analysis of all possible single-nucleotide bulges at each position confirmed that the pivot sites to miR-1, let-7a, miR-155, and miR-7 conferred no better binding than the canonical 6mer-A1 site nested within them. Thus, our results supported the pivot sites proposed for two of the six miRNAs but called into question the generality of this noncanonical site type. Moreover, our approach detected binding of other types of bulged sites, each with a specific bulged nucleotide at target nucleotides 5, 6, 7, or 8, depending on the miRNA (Figure S4). Bulged nucleotides within the miRNA strand abrogated binding, presumably due to steric constraints imposed by AGO.

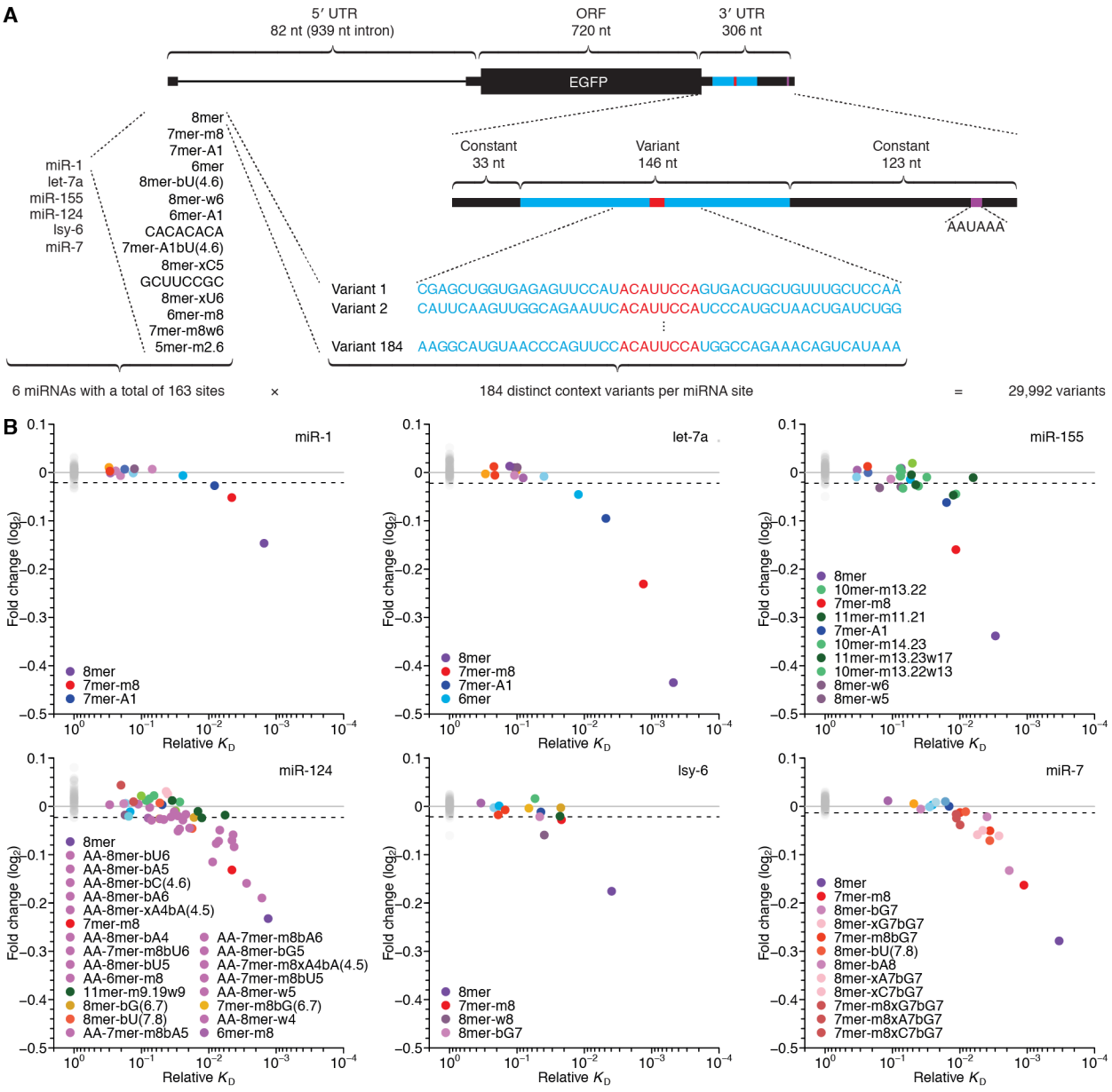


Figure S5. Massively parallel reporter assay to monitor the effects of sites identified by AGO-RBNS. (**A**) Schematic of the EGFP pre-mRNA expressed upon transfection of the library of reporter plasmids. The top, middle, and bottom diagrams respectively depict the pre-mRNA, the 3′ UTR, and a region within the 3′ UTR containing the miR-1 8mer site (red) and its flanking nucleotides (blue). The 163 sites queried corresponded to an earlier list of sites (McGeary et al., 2018), which differed slightly from the current list because it was not informed by the additional AGO-RBNS replicates performed for miR-1, miR-124, and miR-7. (**B**) The relationship between reporter repression efficacy and relative K_D values for all of the queried sites. The relative K_D values are those that were determined when the sites were initially identified (McGeary et al., 2018). When queried in the context of its cognate miRNA, the fold-change ($\log_2$) value of a site was determined by comparing the sum of the counts of all 184 variants corresponding to that site to the average summed counts for these variants observed in the other five transfection experiments (colored points). When queried in the context of each noncognate miRNA, the fold-

change ($\log_2$) value of a site was determined by comparing to the average summed counts from the four other noncognate miRNA transfection experiments (gray points). Each legend lists the sites that mediated repression exceeding twice the standard deviation of the fold-change ($\log_2$) values observed for all the sites not targeted by the transfected miRNA (dashed line).

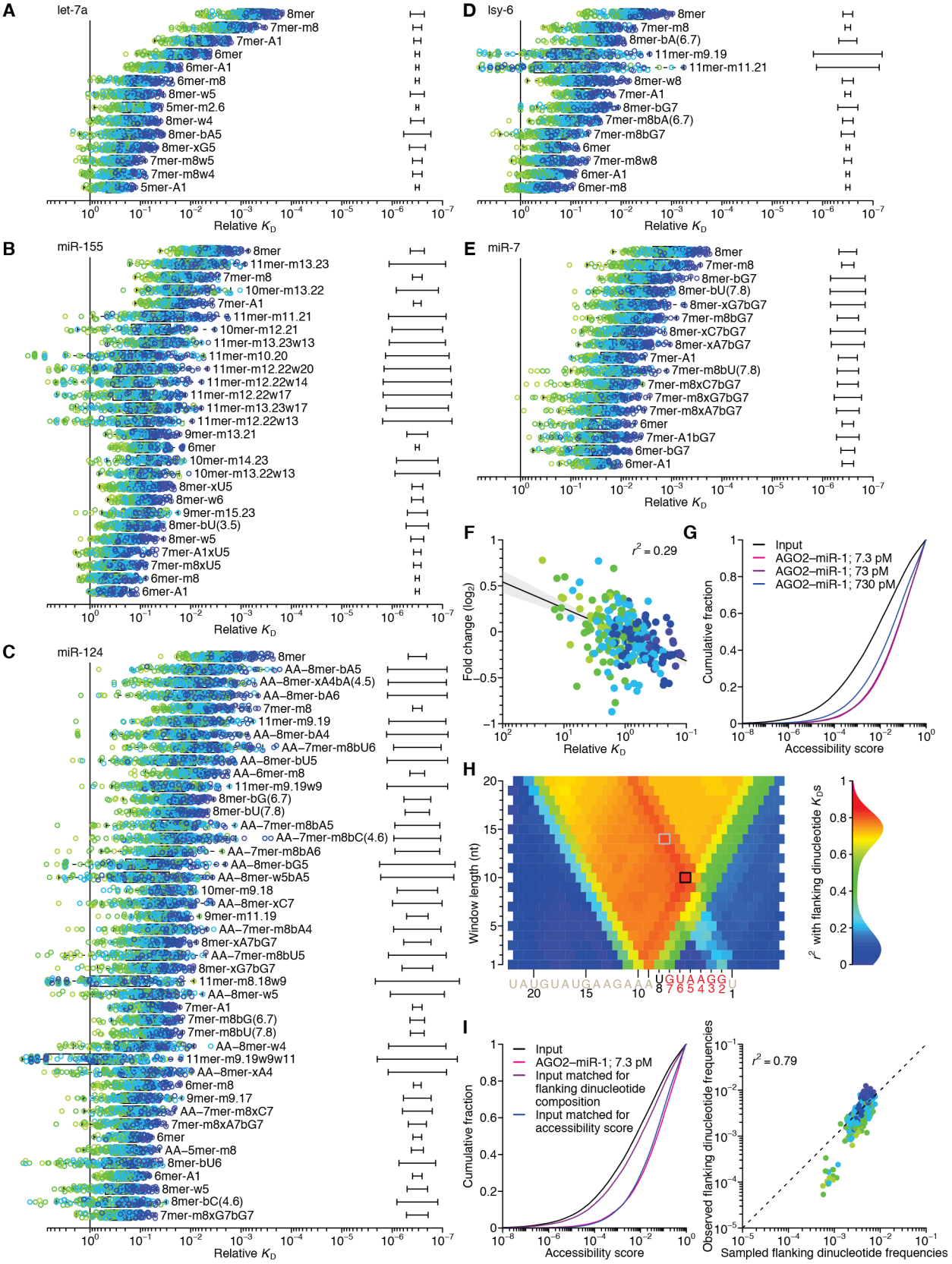

A
let-7a
8mer
7mer-m8
7mer-A1
6mer
6mer-A1
6mer-m8
8mer-w5
5mer-m2.6
8mer-w4
8mer-bA5
8mer-xG5
7mer-m8w5
7mer-m8w4
5mer-A1
Relative K_D
B
miR-155
8mer
11mer-m13.23
7mer-m8
10mer-m13.22
7mer-A1
11mer-m11.21
10mer-m12.21
11mer-m13.23w13
11mer-m10.20
11mer-m12.22w20
11mer-m12.22w14
11mer-m12.22w17
11mer-m13.23w17
11mer-m12.22w13
9mer-m13.21
6mer
10mer-m14.23
10mer-m13.22w13
8mer-xU5
8mer-w6
9mer-m15.23
8mer-bU(3.5)
8mer-w5
7mer-A1xU5
7mer-m8xU5
6mer-m8
6mer-A1
Relative K_D
C
miR-124
8mer
AA–8mer-bA5
AA–8mer-xA4bA(4.5)
AA–8mer-bA6
7mer-m8
11mer-m9.19
AA–8mer-bA4
AA–7mer-m8bU6
AA–8mer-bU5
AA–6mer-m8
11mer-m9.19w9
8mer-bG(6.7)
8mer-bU(7.8)
AA–7mer-m8bA5
AA–7mer-m8bC(4.6)
AA–7mer-m8bA6
AA–8mer-bG5
AA–8mer-w5bA5
10mer-m9.18
AA–8mer-xC7
9mer-m11.19
AA–7mer-m8bA4
8mer-xA7bG7
AA–7mer-m8bU5
8mer-xG7bG7
11mer-m8.18w9
AA–8mer-w5
7mer-A1
7mer-m8bG(6.7)
7mer-m8bU(7.8)
AA–8mer-w4
11mer-m9.19w9w11
AA–8mer-xA4
6mer-m8
9mer-m9.17
AA–7mer-m8xC7
7mer-m8xA7bG7
6mer
AA–5mer-m8
8mer-bU6
6mer-A1
8mer-w5
8mer-bC(4.6)
7mer-m8xG7bG7
Relative K_D
D
lsy-6
8mer
7mer-m8
8mer-bA(6.7)
11mer-m9.19
11mer-m11.21
8mer-w8
7mer-A1
8mer-bG7
7mer-m8bA(6.7)
7mer-m8bG7
6mer
7mer-m8w8
6mer-A1
6mer-m8
Relative K_D
E
miR-7
8mer
7mer-m8
8mer-bG7
8mer-bU(7.8)
8mer-xG7bG7
7mer-m8bG7
8mer-xC7bG7
8mer-xA7bG7
7mer-A1
7mer-m8bU(7.8)
7mer-m8xC7bG7
7mer-m8xG7bG7
7mer-m8xA7bG7
6mer
7mer-A1bG7
6mer-bG7
6mer-A1
Relative K_D
F
r^2 = 0.29
Fold change ($\log_2$)
Relative K_D
G
Input
AGO2–miR-1; 7.3 pM
AGO2–miR-1; 73 pM
AGO2–miR-1; 730 pM
Cumulative fraction
Accessibility score
H
Window length (nt)
UAUGUAUGAAGAAAUGUAAGGU
20 15 10 8 7 6 5 4 3 2 1
r^2 with flanking dinucleotide K_Ds
I
Input
AGO2–miR-1; 7.3 pM
Input matched for flanking dinucleotide composition
Input matched for accessibility score
Cumulative fraction
Accessibility score
r^2 = 0.79
Observed flanking dinucleotide frequencies
Sampled flanking dinucleotide frequencies

Figure S6. The influence of flanking dinucleotide context. (**A**–**E**) Relative K_D values for each flanking dinucleotide combination for each site identified for let-7a (A), miR-155 (B), miR-124 (C), lsy-6 (D), and miR-7 (E). Otherwise, as in Figure 4B. For the larger sites (e.g., the 11-nt 3′-only sites of miR-155, miR-124, and lsy-6), subdividing the low numbers of reads into 144 to 256 categories based on flanking dinucleotide identity resulted in much wider confidence intervals for their respective relative K_D values, and for some pairs of flanking dinucleotides, the number of reads in the input library were too low to estimate a K_D value. (**F**) The relationship between repression efficacy and relative K_D for the 256 flanking dinucleotide combinations. The *x*-axis values are from the linear model in Figure 4C, and the *y*-axis values are from the repression observed in cells, after using a multiple linear regression to distinguish the effect of flanking dinucleotides from that of site type (focusing on repression mediated by 8mer, 7mer-m8, and 7mer-A1 sites). The line shows the best fit to the data (gray region, 95% confidence interval of the trend), determined by least-squares regression weighting residuals using the 95% confidence intervals of the log fold-change estimates. The r^2 value was calculated using similarly weighted Pearson correlation ($p = 5.6 \times 10^{-20}$). The fitted slope of the relationship between fold change ($\log_2$) and relative K_D ($\log_{10}$) for flanking dinucleotide context (0.28 ± 0.06) was in strong agreement with that of the six miRNA site relationships in Figures 3D–3I (mean value of 0.26). (**G**) The cumulative distributions of structural accessibility scores for miR-1 8mer sites in the input (black), the 7.3 pM AGO2–miR-1 (pink), the 73 pM AGO2–miR-1 (purple) and the 730 pM AGO2–miR-1 (blue) libraries. The geometric mean corresponding to each of the four distributions is 2.3×10^{-3}, 2.5×10^{-2}, 2.4×10^{-2}, and 1.3×10^{-2}, respectively. (**H**) The correspondence between relative K_D values for all 256 miR-1 8mer flanking dinucleotide combinations and the geometric mean of the predicted structural-accessibility scores observed for corresponding reads in the input library, as a function of both the length and the position of the sequence segment used for calculating site accessibility. Previous analysis of miRNA targeting indicates that a 14-nt window opposite miRNA positions 1–14 is optimal for calculating the structural-accessibility score, which agrees with an earlier analysis of siRNA efficacy (Agarwal et al., 2015; Tafer et al., 2008). Our analysis also showed that this 14-nt window worked well (gray box, $r^2 = 0.82$), with performance approaching that of the optimum, which was a 10-nt window opposite miRNA positions 1–10 (black box, $r^2 = 0.84$). (**I**) The influence of site accessibility after accounting for nucleotide sequence composition of flanking dinucleotides. Plotted are cumulative distributions of structural-accessibility scores of the 8mer sites of the input library (black), 8mer sites of the bound library from the 7.3 nM sample (red), 8mer sites of the input library from reads sampled to match the accessibility scores of 8mers of the bound library (blue), and 8mer sites of the input library from reads sampled to match the flanking dinucleotide composition of 8mers of the bound library (purple). The geometric mean of the distribution when sampling to match the flanking dinucleotide composition of 8mers of the bound library spanned 21.6% of the difference in geometric means observed between the bound-library and input-library experimental distributions. At the right are the frequencies of dinucleotide combinations flanking miR-1 8mer sites observed in the 7.3 pM AGO2–miR-1 library (left, red line) plotted as a function of the frequencies observed among input reads sampled to match the structural accessibility scores of the reads in the 7.3 pM AGO2–miR-1 library (left, blue line). The r^2 was calculated from the Pearson correlation of log-transformed mean values.

Although we cannot rule out the possibility that the flanking dinucleotide preferences were caused by direct contacts to AGO with sequence preferences that happened to correlate

strongly with those of predicted structural accessibility, the high correspondence of predicted site accessibility and relative K_D—one being the averaged result of a computational algorithm applied to reads from the input library, the other being a biochemical constant derived from AGO-RBNS analyses—strongly implied that site accessibility was the primary cause of the different binding affinities associated with flanking-dinucleotide context (Figures 4D and S6H). Supporting this interpretation, we found that when the 8mer-containing reads of the input library were sampled to match the flanking dinucleotide distribution of the 8mer-containing reads in the 7.3 pM AGO2–miR-1 library, flanking dinucleotide identities explained only a minor fraction of the enrichment of structurally accessible reads observed in the bound libraries (Figure S6I, left). Extending the analysis to data from the other four AGO2–miR-1 concentrations yielded consistent results, with the results from matched sampling of flanking dinucleotides never explaining >25% of the increased mean accessibility score. By contrast, sampling 8mer-containing reads from the input to match the accessibility scores of the bound reads yielded flanking dinucleotide preferences that corresponded to those of the bound library (Figure S6I, right).

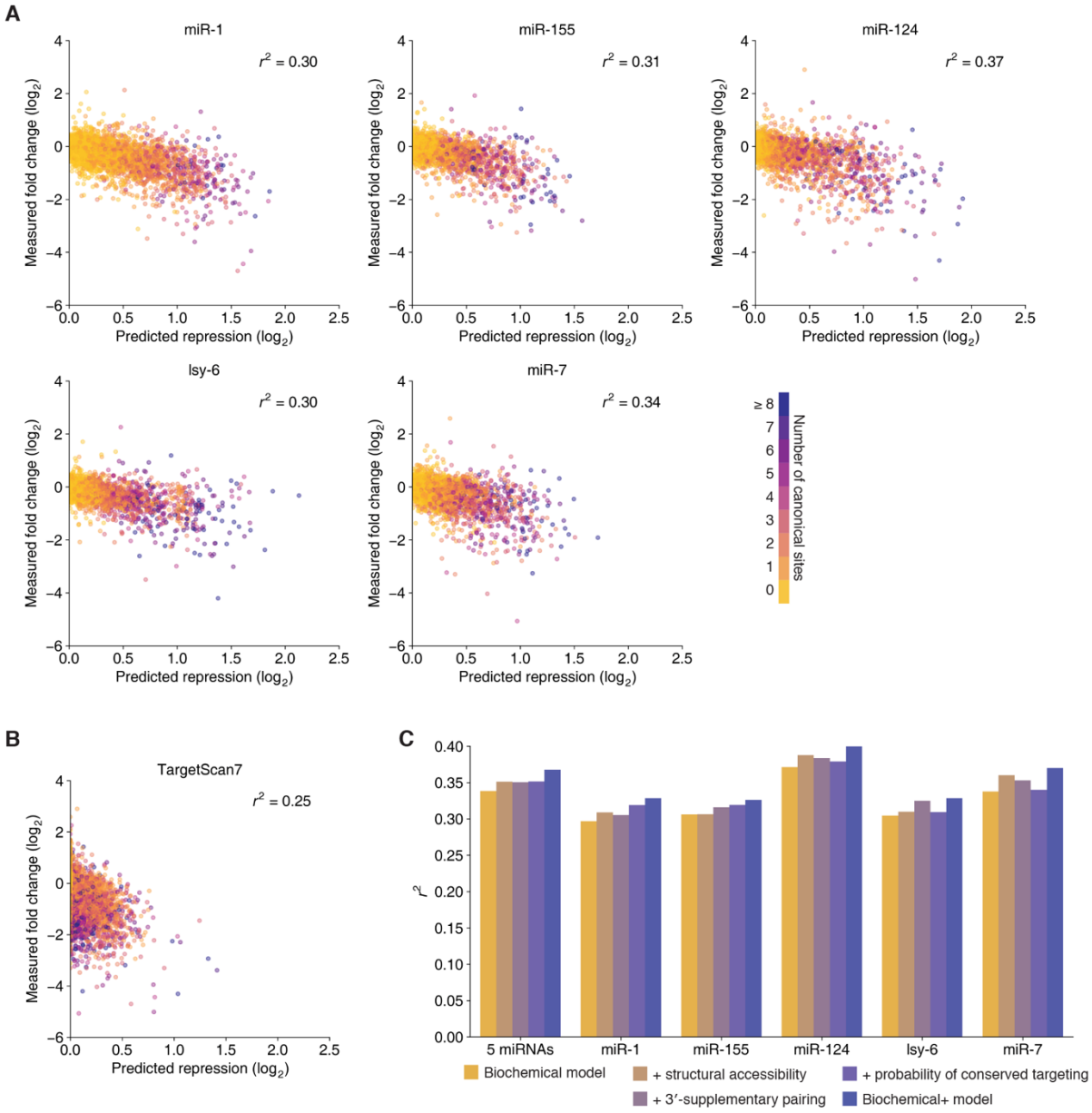


Figure S7. **Additional analyses of the biochemical models.** (**A**) Performance of the biochemical model as evaluated for each of the five miRNAs individually. Otherwise, as in Figure 5C. (**B**) Performance of the published version of the TargetScan7 model as evaluated using the combined results of five miRNAs. Otherwise as in (A). (**C**) Performances of the biochemical model, the biochemical+ model, and three intermediate models as evaluated using the results of the five miRNAs, both in combination (5 miRNAs) and individually. For each of the three intermediate models, a single extra feature of the biochemical+ model (either structural accessibility, 3′-pairing score, or probability of conserved targeting) was incorporated into the biochemical model.

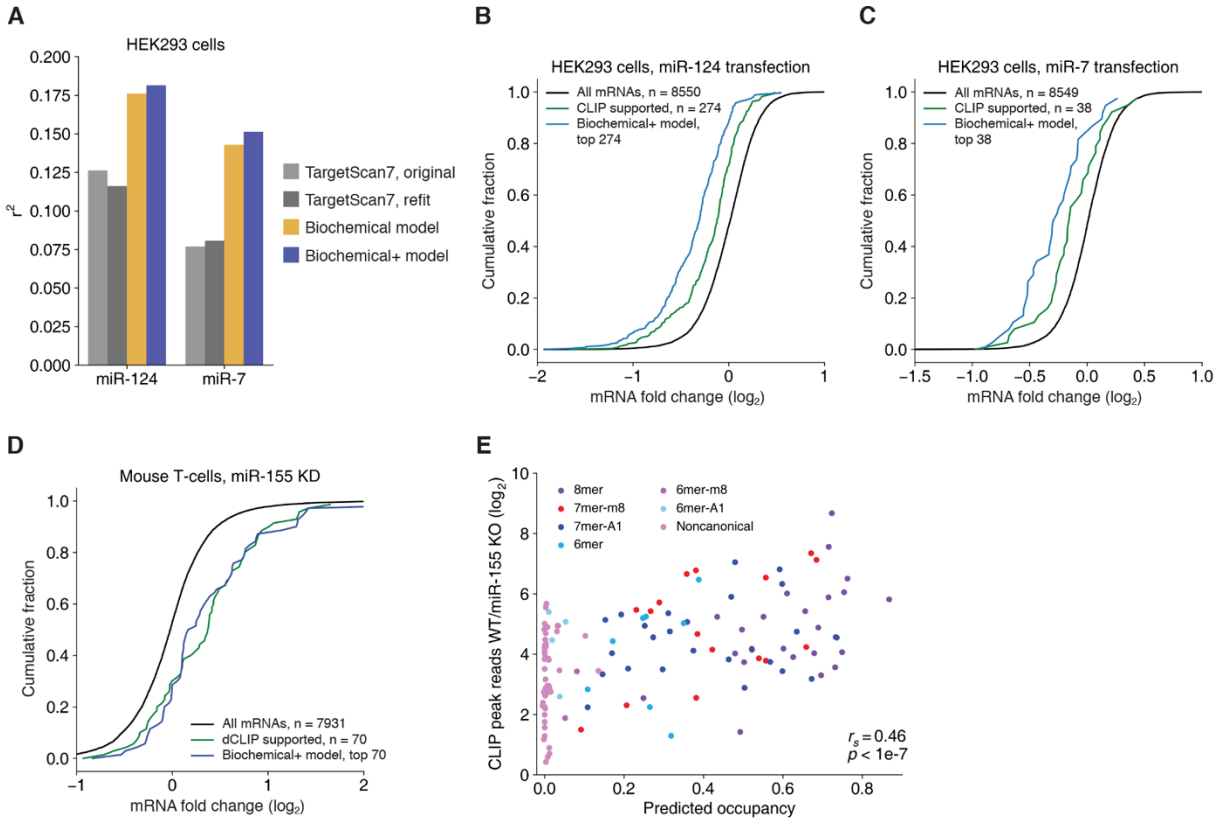


Figure S8. Evaluation of the biochemical models using other published datasets. (**A**) Performances of the biochemical and biochemical+ models compared to those of both the published and refit versions of TargetScan7, as evaluated using mRNA fold changes observed after transfecting either miR-124 or miR-7 into HEK293 cells (Hausser et al., 2009). (**B** and **C**) The ability of the biochemical+ model to identify mRNAs highly responsive to miRNA transfection, compared to that of high-throughput in vivo crosslinking. Plotted are cumulative distributions of mRNA fold changes observed after transfection of either miR-124 (B) or miR-7 (C) into HEK293 cells (Hausser et al., 2009), comparing results for the top targets identified by photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) (Hafner et al., 2010) upon transfection of the cognate miRNA (green) to the results for same number of top targets predicted by the biochemical+ model (blue) and those of all mRNAs (black). (**D**) The ability of the biochemical+ model to identify mRNAs highly responsive to miRNA knockout, compared to that of high-throughput in vivo crosslinking. Results for top targets predicted by the biochemical+ model are compared to those of targets identified by differential CLIP upon knockout of miR-155 in mouse T cells (Loeb et al., 2012). Otherwise as in (B). (**E**) Relationship between enrichment of reads observed at differential CLIP peaks (comparing reads in wild-type to those in miR-155–knockout T cells) and the occupancy of AGO–miR-155 on these CLIP-supported sites as predicted by the biochemical+ model. The Spearman correlation coefficient and *p*-value for this relationship are reported in the bottom right. Points are colored by the identity of the best canonical site type in each CLIP-peak sequence. This relationship was observed for only this CLIP dataset, which was the highest-quality CLIP dataset we evaluated; it had 12 replicates and was the only one that could match the biochemical+ model in identifying top targets (D).

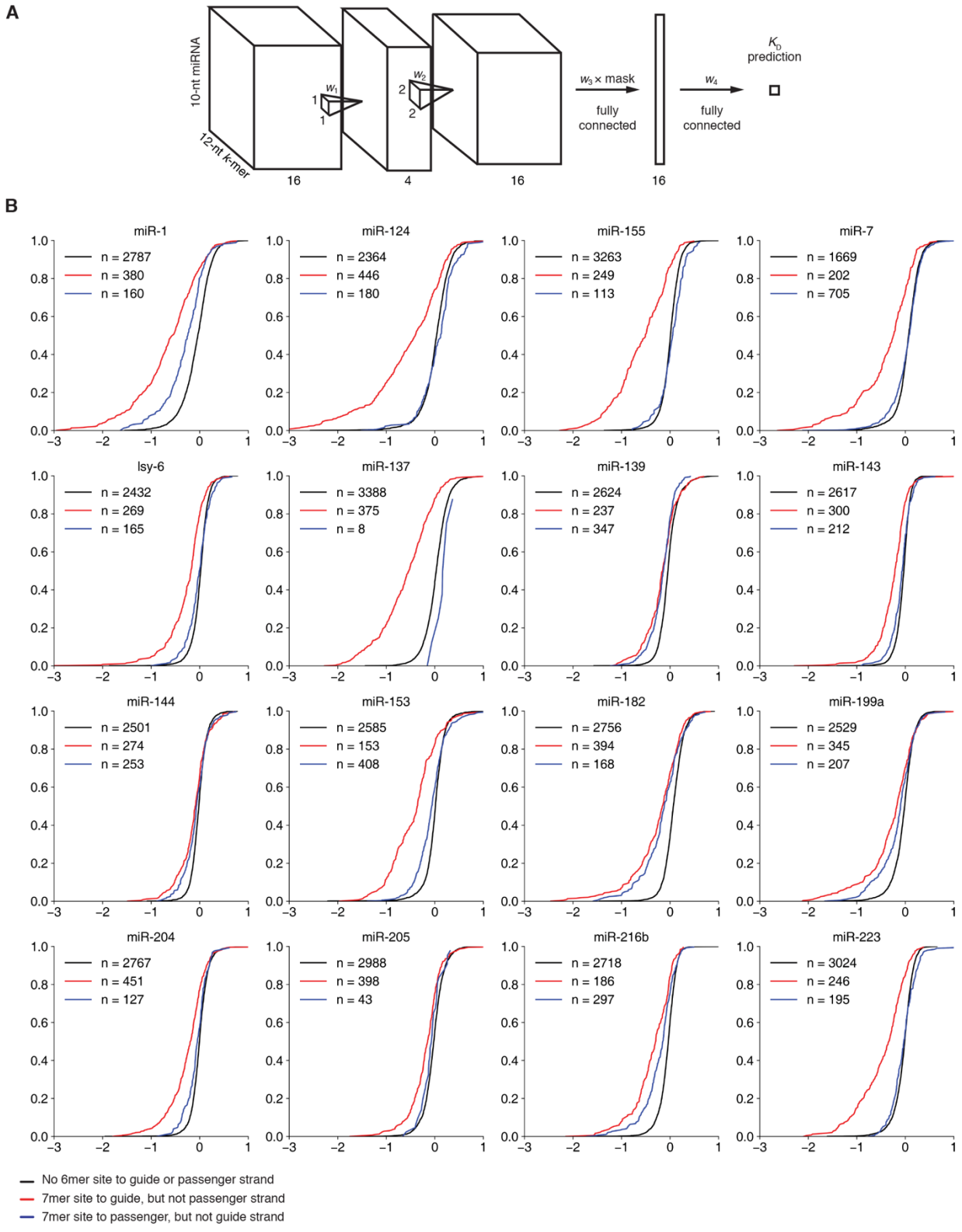


Figure S9. Additional analyses and data related to training the CNN. (**A**) Schematic of the CNN architecture. Each miRNA and 12-nt *k*-mer pair was represented by a 10 × 12 × 16 matrix, where [*i*, *j*, 1 : 16] represented the one-hot encoding of the *i*th nucleotide of the miRNA and the

*j*th nucleotide of the 12-nt *k*-mer. This input was passed through a 1 × 1 convolution with 4 neurons, followed by batch normalization and leaky ReLU activation. This fed into a 2 × 2 convolutional layer with 16 neurons, batch normalization, and leaky ReLU. The third layer was a fully connected layer with 16 neurons, batch normalization, and leaky ReLU. Its weights were multiplied by a mask that preserved weights along the diagonal of miRNA–target pairing, allowing up to 4 nt of offset, and set the remaining weights to 0. The output of this third layer fed into a final fully connected layer to produce the predicted relative K_D value. (**B**) Response of mRNAs to transfected miRNAs used for training. Each plot shows the cumulative distributions of fold-change values in HeLa cells. Results are shown for mRNAs with either a 7–8-nt canonical 3′-UTR site to the transfected miRNA strand (red), a 78-nt canonical 3′-UTR site to the transfected passenger strand (blue), or no canonical site (6mer, 7mer-A1, 7mer-m8, or 8mer) to either strand (black).

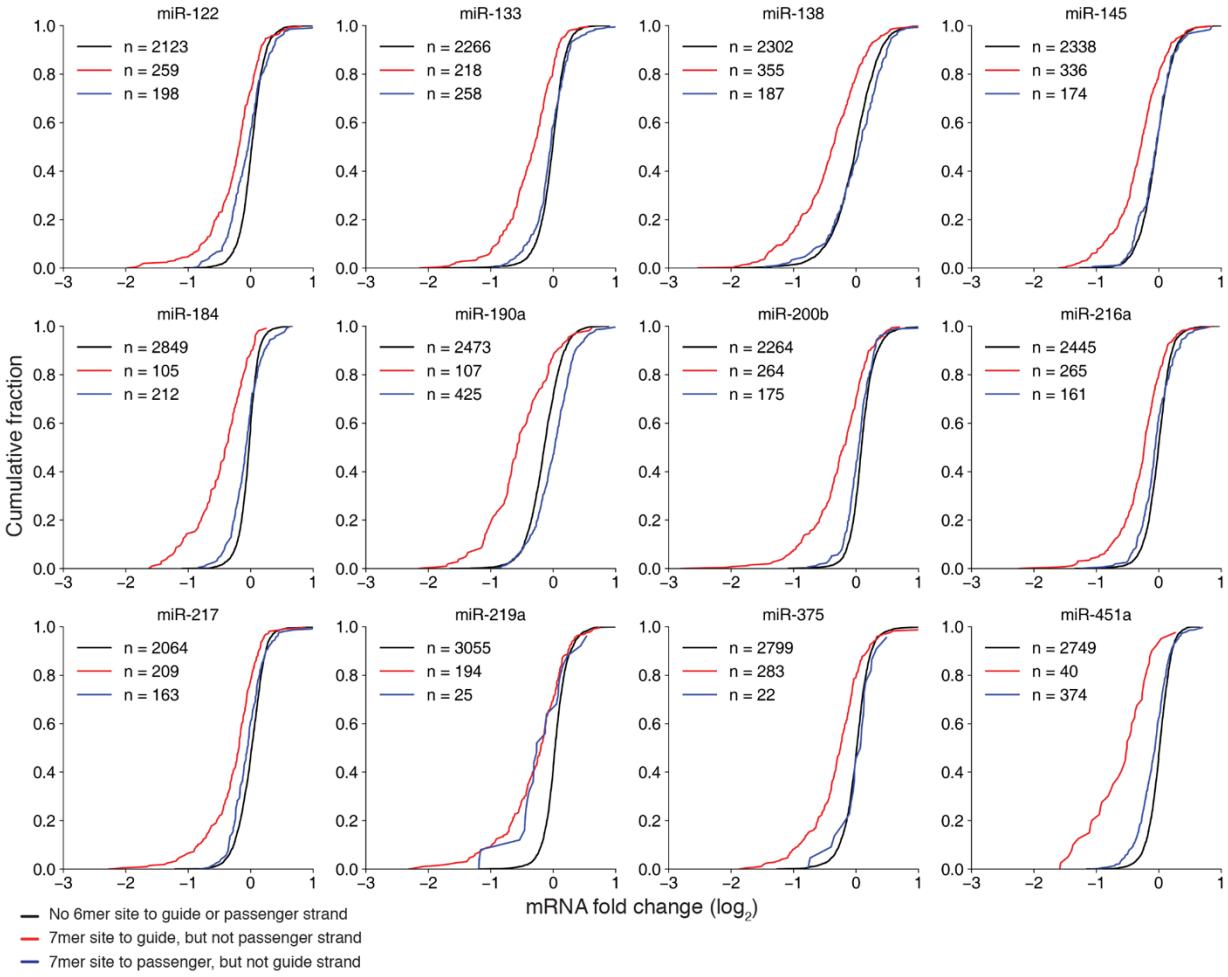


Figure S10. Response of mRNAs to transfected miRNAs used for testing. Each plot shows cumulative distributions of fold-change values of mRNAs in HEK293FT cells. Otherwise, as in Figure S9B.

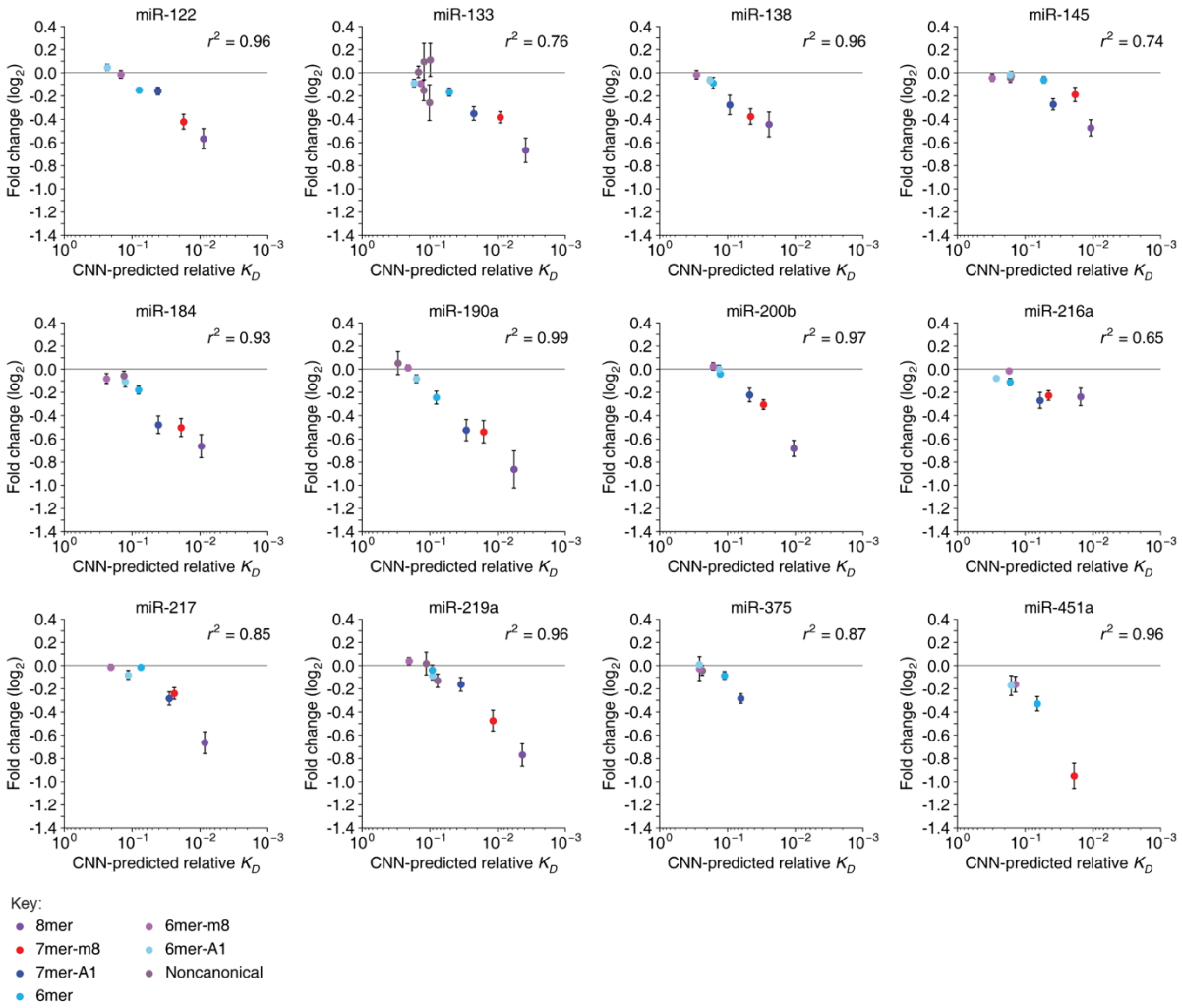


Figure S11. Relationship between mean fold change conferred by each site type in HEK293FT cells and CNN-predicted relative K_D values. Results are shown for the six canonical site types and the predicted noncanonical sites found by examining the 12-nt *k*-mers that had the highest-affinity CNN-predicted K_D values but lacked a canonical site. The miRNAs of the final two panels, miR-375 and miR-451a, contained CpG dinucleotides in their seed regions, which substantially reduced their site abundances in the transcriptome. As a result, the 8mer and 7mer-m8 sites for miR-375 and the 8mer and 7mer-A1 sites for miR-451a each had <20 instances in the 3′ UTRs under consideration, which fell below our threshold for inclusion in this type of analysis, despite these sites having high predicted binding affinities. Otherwise, as in Figures 3D–3I.

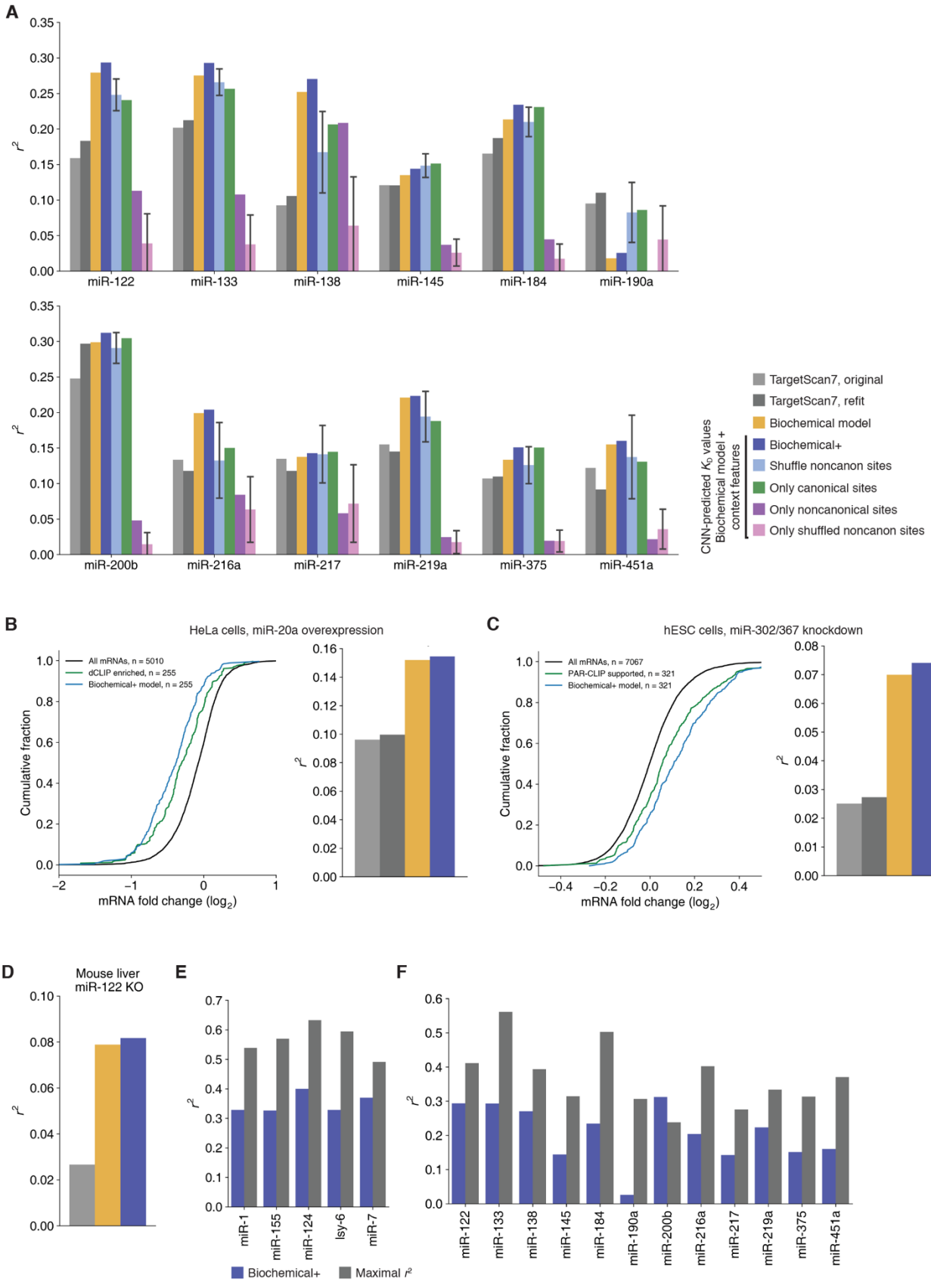

A
B
HeLa cells, miR-20a overexpression
All mRNAs, n = 5010
dCLIP enriched, n = 255
Biochemical+ model, n = 255
Cumulative fraction
mRNA fold change (log2)
C
hESC cells, miR-302/367 knockdown
All mRNAs, n = 7067
PAR-CLIP supported, n = 321
Biochemical+ model, n = 321
D
Mouse liver
miR-122 KO
E
F
miR-122
miR-133
miR-138
miR-145
miR-184
miR-190a
miR-200b
miR-216a
miR-217
miR-219a
miR-375
miR-451a
miR-1
miR-155
miR-124
lsy-6
miR-7
Biochemical+
Maximal r2
TargetScan7, original
TargetScan7, refit
Biochemical model
Biochemical+
Shuffle noncanon sites
Only canonical sites
Only noncanonical sites
Only shuffled noncanon sites
CNN-predicted KD values
Biochemical model +
context features

Figure S12. Additional evaluation of the biochemical models using CNN-predicted K_D values. (**A**) Performance of the models and the contribution of cognate noncanonical sites to performance of the biochemical+ model. Results are shown for each of the 12 miRNAs of the test set used in Figure 6. Otherwise, as in Figure 6D. (**B**) Performance of the biochemical+ model using CNN-predicted K_D values compared to that of differential CLIP (left) and TargetScan (right), as evaluated using mRNA changes observed upon overexpression of miR-20a in HeLa cells (Zhang et al., 2018). Otherwise, as in Figures S8A and S8B. (**C**) Performance of the biochemical+ model using CNN-predicted K_D values compared to that of differential PAR-CLIP (left) and TargetScan (right), as evaluated using mRNA changes observed upon knockdown of miR-302/367 in hESC cells (Lipchina et al., 2011). Otherwise as in (B). (**D**) Performance of the biochemical and biochemical+ models using CNN-predicted K_D values compared to that of TargetScan7, as evaluated using mRNA fold changes observed upon miR-122 knockout in mouse liver cells (Eichhorn et al., 2014). Otherwise, as in Figure S8A. (**E**) Performance of the biochemical+ model (blue) compared with estimated maximal r^2 values (grey) for each of the five miRNAs in Figure 5C. (**F**) Performance of the biochemical+ model using CNN-predicted relative K_D values compared with estimated maximal r^2 values for each of the 12 test miRNAs in Figure 6. Otherwise, as in (E).

Table S1.
Coefficients of linear effects in model of miRNA, site, and flanking-dinucleotide sequence contribution to site binding affinity; related to Figure 4D. The four flanking dinucleotide positions are labeled 5p1, 5p2, 3p1, and 3p2, in the 5′-to-3′ direction (e.g., 5′-$N_{5p1}N_{5p2}$ACAUUCCA$N_{3p1}N_{3p2}$-3′ for the flanking dinucleotide context of the miR-1 8mer site).

	$\Delta\ln(K_D)$		
	Value	Lower CI (2.5%)	Upper CI (97.5%)
miRNA coefficients			
miR-1	−7.30	−7.39	−7.21
let-7a	−8.36	−8.45	−8.27
miR-155	−6.52	−6.61	−6.43
miR-124	−7.22	−7.31	−7.13
lsy-6	−6.16	−6.25	−6.07
miR-7	−7.99	−8.08	−7.90
Site coefficients (with 8mer = 0)			
7mer-m8	0.94	0.85	1.03
7mer-A1	1.55	1.46	1.64
6mer	2.44	2.34	2.54
6mer-m8	5.37	5.28	5.46
6mer-A1	4.45	4.36	4.54
5p1 coefficients (with A = 0)			
C	0.57	0.50	0.63
G	0.86	0.80	0.93
U	0.16	0.10	0.23
5p2 coefficients (with A = 0)			
C	0.62	0.56	0.69
G	1.09	1.03	1.16
U	−0.10	−0.16	−0.04
3p1 coefficients (with A = 0)			
C	0.17	0.10	0.24
G	0.52	0.45	0.59
U	−0.17	−0.24	−0.10
3p2 coefficients (with A = 0)			
C	0.07	−0.01	0.14
G	0.59	0.52	0.67
U	−0.01	−0.09	0.06

Table S2.
Coefficients of pairwise interaction terms of the model described in Table S1 and Figure 4D.

	$\Delta\ln(K_D)$		
	Value	Lower CI (2.5%)	Upper CI (97.5%)
miRNA × site coefficients (with all miRNA × 8mer and all miR-1 × site pairs = 0)			
let-7a × 7mer-m8	0.02	−0.10	0.15
miR-155 × 7mer-m8	0.30	0.17	0.42
miR-124 × 7mer-m8	0.04	−0.08	0.17
lsy-6 × 7mer-m8	0.64	0.52	0.77
miR-7 × 7mer-m8	−0.13	−0.25	−0.00
let-7a × 7mer-A1	0.61	0.49	0.74
miR-155 × 7mer-A1	−0.18	−0.31	−0.06
miR-124 × 7mer-A1	2.04	1.91	2.16
lsy-6 × 7mer-A1	0.73	0.59	0.86
miR-7 × 7mer-A1	1.34	1.21	1.46
let-7a × 6mer	0.63	0.50	0.77
miR-155 × 6mer	0.19	0.06	0.33
miR-124 × 6mer	2.13	1.99	2.27
lsy-6 × 6mer	1.20	1.05	1.35
miR-7 × 6mer	1.23	1.09	1.37
let-7a × 6mer-m8	−0.26	−0.38	−0.13
miR-155 × 6mer-m8	−0.93	−1.06	−0.81
miR-124 × 6mer-m8	−1.68	−1.81	−1.55
lsy-6 × 6mer-m8	−1.14	−1.26	−1.01
miR-7 × 6mer-m8	0.17	0.04	0.29
let-7a × 6mer-A1	−0.39	−0.52	−0.26
miR-155 × 6mer-A1	0.21	0.08	0.33
miR-124 × 6mer-A1	−0.09	−0.22	0.04
lsy-6 × 6mer-A1	−0.80	−0.92	−0.67
miR-7 × 6mer-A1	−1.09	−1.21	−0.96
5p1 × 5p2 coefficients (with all A × N and to N × A = 0)			
C × C	−0.09	−0.18	−0.00
G × C	−0.10	−0.19	−0.01
U × C	0.06	−0.03	0.14
C × G	−0.02	−0.11	0.07
G × G	0.42	0.33	0.52
U × G	0.01	−0.08	0.10
C × U	0.45	0.36	0.54
G × U	0.21	0.11	0.30
U × U	0.29	0.20	0.38
3p1 × 3p2 coefficients (with all A × N and to N × A = 0)			
C × C	0.15	0.05	0.24
G × C	−0.11	−0.21	−0.02
U × C	0.11	0.01	0.20
C × G	−0.11	−0.21	−0.01
G × G	−0.13	−0.23	−0.04
U × G	0.01	−0.09	0.10
C × U	0.07	−0.03	0.17
G × U	−0.03	−0.13	0.06
U × U	−0.03	−0.12	0.07

References

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, e05005.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Ameres, S.L., Martinez, J., and Schroeder, R. (2007). Molecular basis for target RNA recognition and cleavage by human RISC. Cell *130*, 101–112.

Arvey, A., Larsson, E., Sander, C., Leslie, C.S., and Marks, D.S. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. Mol. Syst. Biol. *6*, 363.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bartel, D.P. (2018). Metazoan microRNAs. Cell *173*, 20–51.

Becker, W.R., Ober-Reynolds, B., Jouravleva, K., Jolly, S.M., Zamore, P.D., and Greenleaf, W.J. (2019). High-throughput analysis reveals rules for target RNA binding and cleavage by AGO2. Mol. Cell *75*, 741–755.e11.

Bosson, A.D., Zamudio, J.R., and Sharp, P.A. (2014). Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. Mol. Cell *56*, 347–359.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA–target recognition. PLOS Biol. *3*, e85.

Brown, K.M., Chu, C., and Rana, T.M. (2005). Target accessibility dictates the potency of human RISC. Nat. Struct. Mol. Biol. *12*, 469–470.

Chandradoss, S.D., Schirle, N.T., Szczepaniak, M., Macrae, I.J., and Joo, C. (2015). A dynamic search process underlies microRNA targeting. Cell *162*, 96–107.

Chen, K., Maaskola, J., Siegal, M.L., and Rajewsky, N. (2009). Reexamining microRNA site accessibility in Drosophila: a population genomics study. PLOS ONE *4*, e5681.

Chi, S.W., Hannon, G.J., and Darnell, R.B. (2012). An alternative mode of microRNA target recognition. Nat. Struct. Mol. Biol. *19*, 321–327.

Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S., and Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. Genome Res. *27*, 2015–2024.

Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Mol. Cell *54*, 766–776.

Denzler, R., McGeary, S.E., Title, A.C., Agarwal, V., Bartel, D.P., and Stoffel, M. (2016). Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. Mol. Cell *64*, 565–579.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res. *11*, 1475–1489.

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Nostrand, E.L.V., Pratt, G.A., et al. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Mol. Cell *70*, 854–867.e9.

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

Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2013). Rapid and specific purification of Argonaute-small RNA complexes from crude cell lysates. RNA *19*, 271–279.

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

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs. Nat. Struct. Mol. Biol. *18*, 1139–1146.

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

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous identification of miRNA:target site interactions by different types of ligation reactions. Mol. Cell *54*, 1042–1054.

Gumienny, R., and Zavolan, M. (2015). Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. Nucleic Acids Res. *43*, 1380–1391.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.

Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., and Zavolan, M. (2009). Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C–miRNA complexes and the degradation of miRNA targets. Genome Res. *19*, 2009–2020.

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

Jaganathan, K., Panagiotopoulou, S.K., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting splicing from primary sequence with deep learning. Cell *176*, 535–548.e24.

Jens, M., and Rajewsky, N. (2014). Competition between target sites of regulators shapes post-transcriptional gene regulation. Nat. Rev. Genet. *16*, 113–126.

Jo, M.H., Shin, S., Jung, S.-R., Kim, E., Song, J.-J., and Hohng, S. (2015). Human Argonaute 2 has diverse reaction pathways on target RNAs. Mol. Cell *59*, 117–124.

Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nat. Rev. Genet. *16*, 421–433.

Kedde, M., Strasser, M.J., Boldajipour, B., Vrielink, J.A.F.O., Slanchev, K., Sage, C. le, Nagel, R., Voorhoeve, P.M., Duijse, J. van, Ørom, U.A., et al. (2007). RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. Cell *131*, 1273–1286.

Kedde, M., Kouwenhove, M. van, Zwart, W., Vrielink, J.A.F.O., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. Nat. Cell Biol. *12*, 1014–1020.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nat. Genet. *39*, 1278–1284.

Khorshid, M., Hausser, J., Zavolan, M., and Nimwegen, E. van (2013). A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nat. Methods *10*, 253–255.

Kim, D., Sung, Y.M., Park, J., Kim, S., Kim, J., Park, J., Ha, H., Bae, J.Y., Kim, S., and Baek, D. (2016). General rules for functional microRNA targeting. Nat. Genet. *48*, 1517–1526.

Klum, S.M., Chandradoss, S.D., Schirle, N.T., Joo, C., and MacRae, I.J. (2018). Helix-7 in Argonaute2 shapes the microRNA seed region for rapid target recognition. EMBO J. *37*, 75–88.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–900.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M., et al. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. Cell *129*, 1401–1414.

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

Linsley, P.S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M.M., Bartz, S.R., Johnson, J.M., Cummins, J.M., Raymond, C.K., Dai, H., et al. (2007). Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. Mol. Cell Biol. *27*, 2240–2252.

Lipchina, I., Elkabetz, Y., Hafner, M., Sheridan, R., Mihailovic, A., Tuschl, T., Sander, C., Studer, L., and Betel, D. (2011). Genome-wide identification of microRNA targets in human ES cells reveals a role for miR-302 in modulating BMP response. Genes Dev. *25*, 2173–2186.

Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., and Rudensky, A.Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol. Cell *48*, 760–770.

Lorenz, R., Bernhart, S.H., Siederdissen, C.H. zu, Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

McGeary, S.E., Lin, K.S., Shi, C.Y., Bisaria, N., and Bartel, D.P. (2018). The biochemical basis of microRNA targeting efficacy. bioRxiv 414763.

Nam, J.-W., Rissland, O.S., Koppstein, D., Abreu-Goodger, C., Jan, C.H., Agarwal, V., Yildirim, M.A., Rodriguez, A., and Bartel, D.P. (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. Mol. Cell *53*, 1031–1043.

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

Obernosterer, G. (2006). Post-transcriptional regulation of microRNA expression. RNA *12*, 1161–1167.

Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., and Hatzigeorgiou, A.G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res. *41*, W169–W173.

R Core Team (2014). R: a language and environment for statistical computing.

Rio, D.C. (2013). Expression and purification of active recombinant T7 RNA polymerase from E. coli. Cold Spring Harb. Protoc. doi:10.1101/pdb.prot078527.

Rudnick, S.I., Swaminathan, J., Sumaroka, M., Liebhaber, S., and Gewirtz, A.M. (2008). Effects of local mRNA structure on posttranscriptional gene silencing. Proc. Natl. Acad. Sci. USA *105*, 13787–13792.

Saito, T., and Sætrom, P. (2012). Target gene expression levels and competition between transfected and endogenous microRNAs are strong confounding factors in microRNA high-throughput experiments. Silence *3*, 3.

Salomon, W.E., Jolly, S.M., Moore, M.J., Zamore, P.D., and Serebrov, V. (2015). Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. Cell *162*, 84–95.

Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. Science *346*, 608–613.

Schirle, N.T., Sheu-Gruttadauria, J., Chandradoss, S.D., Joo, C., and MacRae, I.J. (2015). Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. eLife *4*, e07646.

Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D.S., and Oudenaarden, A. van (2015). MicroRNA control of protein expression noise. Science *348*, 128–132.

Shin, C., Nam, J.-W., Farh, K.K.-H., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. Mol. Cell *38*, 789–802.

Tafer, H., Ameres, S.L., Obernosterer, G., Gebeshuber, C.A., Schroeder, R., Martinez, J., and Hofacker, I.L. (2008). The impact of target site accessibility on the design of effective siRNAs. Nat. Biotechnol. *26*, 578–583.

Tunney, R., McGlincy, N.J., Graham, M.E., Naddaf, N., Pachter, L., and Lareau, L.F. (2018). Accurate design of translational output by a neural network model of ribosome distribution. Nat. Struct. Mol. Biol. *25*, 577–582.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. Cell *151*, 1055–1067.

Zhang, K., Zhang, X., Cai, Z., Zhou, J., Cao, R., Zhao, Y., Chen, Z., Wang, D., Ruan, W., Zhao, Q., et al. (2018). A novel class of microRNA-recognition elements that function only within open reading frames. Nat. Struct. Mol. Biol. *25*, 1019–1027.

Chapter 3.

Pairing to the microRNA 3′ region occurs through two alternative binding modes, with affinity shaped by pairing position and microRNA G nucleotides

Sean E. McGeary[1,2,3]*, Namita Bisaria[1,2,3]*, Kathy S. Lin[1,2,3,4], and David P. Bartel[1,2,3,4]

[1]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*These authors contributed equally to this work.

N.B performed the experiments and the initial analyses, with help from S.E.M. S.E.M. performed all final analyses. K.S.L. provided computational expertise. S.E.M., N.B., and D.P.B. designed the study and wrote the manuscript.

Work in this chapter is in revision.

Abstract

MicroRNAs (miRNAs), in association with Argonuate (AGO) proteins, direct repression by pairing to sites within mRNAs. Compared to pairing preferences of the miRNA seed region (nucleotides 2–8), preferences of the miRNA 3′ region are poorly understood, due to sparsity of measured affinities for the many possibilities. We used RNA bind-n-seq with purified AGO2–miRNA complexes to measure relative affinities of >1,000 3′-pairing architectures. Optimal 3′ pairing compensated for a seed mismatch to increase affinity by up to >500-fold. Some miRNAs had two high-affinity 3′-pairing modes—one of which allowed pairing to miRNA nucleotide 11 but required additional nucleotides to bridge seed and 3′ pairing. Both the affinity of the binding modes and the position of optimal pairing tracked with the occurrence of G or oligo(G) nucleotides within the miRNA. These and other results advance understanding of miRNA targeting, providing insight into how optimal 3′ pairing is determined for each miRNA.

Highlights

• RNA bind-n-seq reveals relative affinities of >1,000 3′-pairing architectures

• Two distinct 3′-binding modes enhance affinity—by as much as >500-fold for some miRNAs

• G and poly(G) nucleotides help define the miRNA 3′ segment most critical for pairing

• Seed mismatch identity can influence the contribution of compensatory 3′ pairing

Introduction

MicroRNAs (miRNAs) are ~22-nt regulatory RNAs that are processed from hairpin precursors. Upon processing, miRNAs associate with an Argonaute (AGO) protein and pair to sites within mRNAs to direct the destabilization and/or translational repression of these mRNA targets (Bartel, 2018; Jonas and Izaurralde, 2015). For most sites that confer repression in mammalian

cells, pairing to miRNA nucleotides 2–7, referred to as the miRNA seed, is critical for target recognition, with an additional pair to miRNA position 8 or an A across from miRNA position 1 often enhancing targeting efficacy (Bartel, 2009; Lewis et al., 2005). Such sites with a perfect 6–8-nucleotide (nt) match to the miRNA seed region (Figure 1A) are heuristically predictive of repression, with longer sites being more effective than shorter ones and more sites being more effective than fewer sites (Agarwal et al., 2015; Grimson et al., 2007). In addition, contextual features extrinsic to a site itself can also influence targeting efficacy. For example, sites are typically more effective if they reside in the 3′ untranslated region (UTR) and out of the path of either the scanning initiation complex or the translating ribosome (Grimson et al., 2007). They are also more effective if they reside either near to other sites that can act cooperatively or within a region that is not predicted to form occlusive secondary structure (Agarwal et al., 2015; Grimson et al., 2007; McGeary et al., 2019; Sætrom et al., 2007; Wan et al., 2014).

Pairing to the miRNA 3′ region, particularly pairing that includes miRNA nucleotides 13–16, can supplement perfect seed pairing to enhance targeting efficacy beyond that of seed pairing alone, and extensive pairing to the 3′ region can compensate for imperfect seed pairing to enable consequential repression (Brennecke et al., 2005; Grimson et al., 2007; Lewis et al., 2005). These two bipartite site types are referred to as 3′-supplementary and 3′-compensatory sites, respectively (Figure 1A). Although 3′-supplementary sites are less common than sites with only a seed match, comprising ~5% of all conserved sites observed in mammals, thousands of sites with preferentially conserved 3′-supplementary pairing are present in human 3′ UTRs (Friedman et al., 2009). Conserved 3′-compensatory sites are even less common, comprising only ~1.5% of all preferentially conserved sites observed in human 3′ UTRs (Friedman et al., 2009). Nonetheless, two instances of this relatively rare site type within the 3′ UTR of *lin-41* mediate the extreme morphological and developmental defects by which the *let-7* miRNA was

discovered in *C. elegans* (Ecsedi et al., 2015; Reinhart et al., 2000). Moreover, the use of these 3′-compensatory sites rather than canonical sites for *lin-41* repression is consequential; site mutations that create perfect pairing to the let-7 seed cause precocious repression of the mRNA by other members of the let-7 seed family expressed during earlier larval stages (Brancati and Großhans, 2018). These results support the notion that 3′-compensatory sites enable differential target specificity between miRNAs that share common seed sequences but differ within their 3′ regions (Brennecke et al., 2005; Lewis et al., 2005).

Although the global analyses of site conservation and efficacy provide compelling evidence that pairing to the miRNA 3′ region is also utilized in mammalian cells, these analyses leave many questions unanswered. For example, analysis of site conservation can provide an estimate of the number of sites with 3′-supplementary/compensatory pairing that are under purifying selection, but among these sites, it cannot cleanly distinguish those under selection from those conserved by chance (Friedman et al., 2009). Likewise, global analysis of the effects of perturbing miRNAs on mRNA levels (or on translational efficiency) is most reliable when averaging the effects over sites from many mRNAs (Grimson et al., 2007), which can obscure the identification of particularly efficacious sites. Global analyses of site conservation and efficacy also become less useful when examining rarer site types. For example, obtaining a reliable signal for preferential conservation of 3′-supplementary/compensatory pairing requires aggregating data from multiple miRNAs, which obscures differences between miRNAs, and even when aggregating multiple miRNA-perturbation (e.g., transfection) datasets, which enables efficacy of 3′-supplementary sites to be detected, a signal for the efficacy of the rarer 3′-compensatory sites has not been detected.

Understanding the contribution of pairing to the miRNA 3′ region is further complicated by the vast number of possible variations in 3′-pairing architecture. When describing the pairing

architecture of a 3′-compensatory site, five characteristics must be specified: 1) the length of contiguous pairing between the site and the miRNA 3′ region, 2) the position of pairing to the miRNA 3′ region, as defined by the 5′-most miRNA nucleotide involved in 3′ pairing, 3) the difference between the number of unpaired target nucleotides and number of unpaired miRNA nucleotides bridging the seed and 3′ pairing, hereafter referred to as the "3′-pairing offset," 4) the nature of the imperfect pairing to the seed, and 5) the nature of any imperfections in the 3′ pairing (Figure 1B). When considering only sites with perfect 3′ pairing with lengths ranging from 4–11 base pairs (bp), offsets ranging from −4 to +16 nt, and seed pairing interrupted by one of 18 possible single mismatches (or wobbles) to the 6-nt seed, there are >16,000 possible variants to the site architecture. For any miRNA under consideration, most of these variants are not present in the transcriptome, which limits the utility of global analyses of conservation or efficacy, or any other approach that requires one or more instance of the site in the transcriptome.

The observation that miRNA targeting efficacy observed in the cell is largely a function of the affinity between the AGO–miRNA complex and the site (McGeary et al., 2019), indicates that contributions of 3′ pairing to affinities measured in vitro can provide insight into biological targeting efficacy. Affinities for the sites that have been measured reveal some differences between miRNAs and a striking effect of longer pairing (Becker et al., 2019; Salomon et al., 2015; Sheu-Gruttadauria et al., 2019a, 2019b; Wee et al., 2012). For example, pairing to positions 13–16 imparts only a 2-fold increase in binding affinity for let-7a (Wee et al., 2012) and miR-122 (Sheu-Gruttadauria et al., 2019b), but an 11-fold increase for miR-21 (Salomon et al., 2015), raising the question of whether these differences are due to different miRNAs having different capacities to benefit from 3′ pairing, or distinct optimal positions or offsets of pairing. Alternatively, these differences might be attributable to the particular non-3′-paired, seed-only sites used for reference. Another report showed that 10 bp of 3′-supplementary pairing could

decrease the dissociation rate constant (k_{off}) of a miR-122 site by 20-fold (with a presumed corresponding increase in affinity), whereas 9 bp of 3′-supplementary pairing (including a terminal G:U wobble) could increase the binding affinity of a miR-27a site by >400-fold (Sheu-Gruttadauria et al., 2019b). In another report, the binding affinity of two synthetic variants of miR-122 was shown to vary ~10-fold with the extent of 3′-pairing offset (Sheu-Gruttadauria et al., 2019a), as examined in the context of one seed site-type (7mer-m8), one 3′-pairing length (4 bp involving miRNA nucleotides 13–16), target RNAs that terminated immediately after pairing to nucleotide 16, and with poly(A) sequence bridging the seed and supplementary pairing. Taken together, these reports unambiguously demonstrate the potential for miRNA 3′ pairing to enable high-affinity binding, and also illustrate that the realized benefit of this pairing varies considerably depending on the miRNA sequence and the particular architecture of the seed and 3′ pairing of the target site. Owing to the large number of such pairing possibilities for even a single miRNA, a precise description of how these features together modulate the benefit of 3′ pairing will be possible only after acquiring many more measurements.

Imaging-based, high-throughput single-molecule biochemistry has recently been applied to acquire affinity measurements for ~23,000 sites for each of two miRNAs (let-7a and miR-21), including many sites with 3′ pairing (Becker et al., 2019). These measurements revealed that miR-21 relies more on 3′ pairing when binding to a fully complementary target than does let-7a, that homopolymeric insertions are the least disruptive to binding when inserted between nucleotides 8 and 11 within the context of fully complementary binding, and that mismatches near the miRNA 3′ terminus (after position 16) increase target slicing and decrease binding affinity. However, because the design of target libraries was based primarily on fully complementary RNA targets to which varying extents of mismatched, bulged, and deleted nucleotides were introduced, only a small minority of the target RNAs queried possess 3′-

compensatory sites, which have both a seed mismatch and intermediate complementarity to the miRNA 3′ end. Furthermore, most of the target RNAs that do possess either a 3′-compensatory site or a 3′-supplementary site have a site with an offset of 0 nt, leading to the exclusion of most potential site architectures. A fuller understanding of the contribution to pairing to the miRNA 3′ region requires the acquisition of many more affinity measurements with target RNA sequences that vary with respect to their seed pairing, and the position, offset, and length of 3′ pairing.

RNA bind-n-seq (RBNS) enables unbiased, high-throughput assessment binding sites embedded within a larger random-sequence context (Dominguez et al., 2018; Lambert et al., 2014). We recently adapted RBNS for the study of miRNA targeting, and we built an analysis pipeline enabling calculation of relative equilibrium dissociation constants (K_D values) for many thousands of different RNA *k*-mers ≤12 nt in length, which allowed for quantitative comparisons of putative site types and sequence features that would not be possible by analysis of *k*-mer enrichment alone (McGeary et al., 2019). Applying this AGO-RBNS platform to AGO–miRNA complexes with six different miRNAs revealed noncanonical target sites specific for each miRNA, miRNA-specific differences in canonical target-site affinities, and large effects of nucleotides flanking each site (McGeary et al., 2019). Here, we further adapted the AGO-RBNS protocol to enable examination of sites >12 nt in length, thereby enabling the high-throughput investigation of bipartite sites containing near-perfect seed pairing and 4–11 additional pairs to the miRNA 3′ region. We applied this modified protocol to the systematic interrogation of the contribution of 3′ pairing for three natural miRNA sequences and four synthetic derivatives.

RBNS measures affinities for many 3′-compensatory sites of let-7a

As previously implemented, AGO-RBNS utilizes a series of binding reactions, each containing an RNA library at a concentration of 100 nM and a purified AGO–miRNA complex at one of

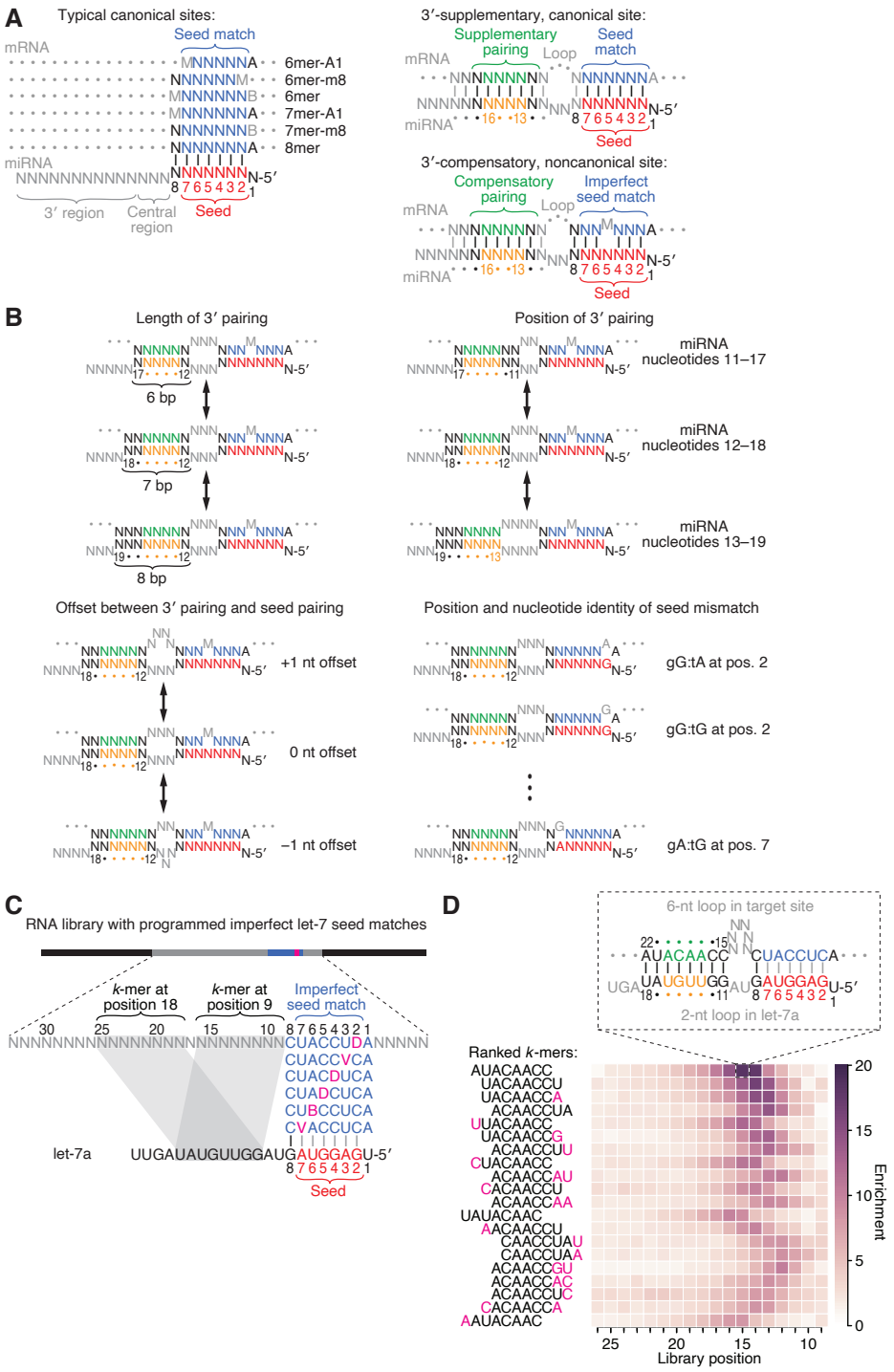


Figure 1. Features of miRNA 3′-compensatory sites characterized using AGO-RBNS. (**A**) Pairing of typical canonical sites (left), 3′-supplementary, canonical sites (upper right), and 3′-compensatory, noncanonical sites (lower right). Canonical sites contain contiguous pairing (blue) to the seed (red). Sites with shifted complementarity (i.e., the 6mer-A1 and 6mer-m8) are sometimes also classified as canonical sites. 3′-supplementary sites have canonical seed pairing in addition to pairing to the miRNA 3′ region, typically including pairing (green) to miRNA nucleotides 13–16 (yellow). 3′-compensatory sites contain fewer than six nucleotides of

contiguous Watson–Crick pairing to the seed region and additional pairing to the 3′ region, also typically including pairing to miRNA nucleotides 13–16. N represents A, C, G, or U; B represents C, G, or U; vertical lines represent Watson–Crick pairing, and M opposite N represents a non-Watson–Crick match. (**B**) Four independent features that define architectures of 3′-compensatory sites with single seed mismatches: 1) the length of 3′ pairing (upper left), measured as the number of contiguous base pairs to the miRNA 3′ region; 2) the position of 3′ pairing (upper right), defined as the 5′-most miRNA nucleotide engaged in 3′ pairing; 3) the offset between the seed pairing and 3′ pairing (lower left), which specifies the number of unpaired nucleotides separating the seed- and 3′-paired segments in the target RNA relative to that in the miRNA; and 4) the position and identity of the mismatch to the seed (lower right). 3′-compensatory architectures can also differ due to mismatched or bulged nucleotides within the 3′ pairing, which is not shown. (**C**) A programmed AGO-RBNS RNA library for let-7a. The library contains an 8-nt region with all 18 possible single-nucleotide mismatches (pink) to the let-7a seed (red), with 25 nt of random-sequence RNA upstream of this region and 5 nt of random-sequence RNA downstream. Library positions are numbered with respect to the programmed 8-nt mismatched site. B represents C, G, or U; D represents A, G, or U; V represents A, C, or G; N represents A, C, G, or U. The black vertical line depicts perfect pairing to position 8, and gray vertical lines indicate a non-Watson–Crick match somewhere within the seed pairing. (**D**) The top 20 8-nt *k*-mers identified by AGO-RBNS performed with the highest concentration of AGO2–let-7a (840 pM) and the programmed library (100 nM). *k*-mers were ranked by the sum of their enrichments at the five positions of the library at which they were most enriched. Left, alignment of *k*-mers, indicating in pink nucleotides that were not Watson–Crick matches to the miRNA. Right, heat map showing *k*-mer enrichment at each position of the library, with pairing shown for the top 8-nt *k*-mer at the position of the library at which it was most enriched. Black vertical lines depict perfect Watson–Crick pairing, and gray vertical lines indicate a non-Watson–Crick match somewhere within the seed pairing.

several concentrations spanning a 100-fold range. Each molecule of the RNA library has a central region of 37 random-sequence nucleotides flanked by constant sequences on each side that enable preparation of sequencing libraries. Upon reaching binding equilibrium, each reaction is passed through a nitrocellulose membrane, which retains the AGO–miRNA complex and any library molecules that are bound to the complex. These bound library molecules are isolated and subjected to high-throughput sequencing, along with the input RNA library. For any *k*-mers ≤12 nt, binding can be detected as enrichment in the bound compared to input sequences. Furthermore, relative K_D values can be estimated simultaneously for hundreds of thousands of

different *k*-mers by fitting a biochemical model to *k*-mer fractional abundances from each of the bound libraries.

As originally implemented, AGO-RBNS cannot provide reliable information on sites with more than ~5 supplementary/compensatory pairs because such sites, which involve >12 nt of total pairing (Figure 1A, right), are too rare in the sequences obtained from input RNA library to enable accurate calculation of enrichment values. To overcome this constraint for sites to let-7a, we replaced the random-sequence library with a library that was heavily enriched in 3′-compensatory sites to let-7a because each molecule of the library was designed to contain a programmed region of imperfect seed pairing to let-7a embedded within the random-sequence region, with 25 and 5 nt of random-sequence RNA separating the programmed region from the 5′ and 3′ constant sequences, respectively (Figure 1C). In each library molecule, this programmed region of imperfect seed pairing matched let-7a at positions 1 and 8, and at all but one position of its 6-nt seed, such that each library molecule contained one of 18 possible single-nucleotide seed mismatches (including wobbles) in approximately equal proportion. With this programmed region of imperfect seed pairing, each library contained 3′-compensatory sites at a ~250-fold greater frequency than expected for a fully randomized RNA library.

AGO-RBNS was performed using this programmed library and purified AGO2–let-7a. For our initial analysis, we calculated the enrichment of all 8-nt *k*-mers at each position between the two constant regions of the library. To survey preferred 3′-pairing positions and offsets, we ranked these *k*-mers on the basis of the enrichment observed at their five most optimal offsets and examined the top 20 *k*-mers 8 nt in length (Figure 1D). The most enriched was AUACAACC—the perfect Watson–Crick match to positions 11–18 of the let-7a miRNA (Figure 1D). This 8-nt 3′ site was most strongly enriched when starting at position 15 of the library, thereby creating an internal loop with two miRNA nucleotides (9 and 10) and six target-site

nucleotides (positions 9–14) separating seed pairing and 3′ pairing (Figure 1D, top). Using our nomenclature (Figure 1B), this 3′ site was classified as a position-11 site with pairing length of 8 bp and offset of +4 nt. This 8-nt position-11 site was also ≥5-fold enriched at seven other neighboring offsets, indicating that looping out 3–10 unpaired library nucleotides opposite miRNA nucleotides 9 and 10 was tolerated, albeit to varying degrees (Figure 1D).

The second-most enriched 8-nt *k*-mer was UACAACCU—the perfect Watson–Crick match to let-7a positions 10–17 (Figure 1D). This 3′ site had a maximal enrichment with 5, rather than 6, unpaired library nucleotides spanning the seed and 3′ pairing, with the distribution of enrichments shifted by 1 nt in comparison to that of the AUACAACC site. This 1-nt shift in the enrichment distribution corresponded with the 1-nt shift in site position (from 11 to 10 of the miRNA) to maintain an optimal offset of +4 target nucleotides. Indeed, the next 18 most enriched 8-nt *k*-mers represented 3′ sites with the pairing positions ranging from miRNA nucleotides 9–12 and enrichment distributions that correspondingly shifted to maintain an optimal offset of +4 target nucleotides (Figure 1D). Each had a contiguous stretch of 6–8 perfect Watson–Crick pairs to the let-7a 3′ region, usually including the ACAACC 6-nt *k*-mer, which suggested that perfect pairing to let-7a positions 11–16, with a +4-nt offset, was particularly important for enhancing site affinity.

let-7a has two distinct 3′-pairing modes

For more comprehensive examination of 3′ sites of varied lengths, positions, and offsets (Figure 1B), we enumerated 3′ sites of lengths 4–11 nt that perfectly paired to the miRNA starting at any position downstream of nucleotide 8. For each length and position of 3′ pairing (e.g., for the 4mer-m9–12, the 4mer-m10–13, etc.), we further enumerated all pairing offsets compatible with the 3′ site residing within the 25-nt random-sequence region upstream of the programmed site,

which resulted in 1006 distinct 3′-site possibilities. For our initial K_D estimation and analyses, we did not distinguish between the 18 possible seed-mismatch types, which increased the reads for each 3′-site possibility, thereby enabling examination of sites as long as 11 nt. We also enumerated each canonical site (including the 6mer-m8 and 6mer-A1 sites, Figure 1A) residing with the 25-nt random-sequence region, as well as each of the 18 single-nucleotide seed-mismatch sites residing within this region.

Simultaneous estimation of the fractional abundance of these sites in each of the AGO2–let-7a-bound libraries in comparison to that of the input library enabled calculation of their relative K_D values. The relative K_D values corresponding to 3′ pairing spanned a >500-fold range (Figure 2A), with strong agreement observed between the results of replicate experiments performed independently with different preparations of both AGO2–let-7a and the let-7a programmed library (r^2 = 0.96, Figure S1A, left). Agreement between the two replicates was maintained when assigning each site to one of 18 3′-compensatory sites, each with a different single-nucleotide seed mismatch (r^2 = 0.78, n = 23,912; Figure S1A, right), albeit to a lesser degree, illustrating the utility of pooling the results for different seed mismatches to obtain higher sequencing coverage when querying each 3′-pairing possibility. Furthermore, for shorter 3′ sites, which could be analyzed using data from a standard AGO-RBNS experiment that used a non-programmed random-sequence library (McGeary et al., 2019), the relative K_D values determined from the programmed library correlated well with those determined from a random-sequence library (r^2 = 0.83, Figure S1B). Despite the overall correlation, a minor systematic difference in the values for the same sites determined from the two types of libraries was observed. This offset was attributable to a distortion caused by the absence of no-site-containing RNA molecules in the programmed library and was corrected accordingly (Figure S1B).

Plotting the cumulative distribution of affinities for the 1006 3′-compensatory sites stratified according to their length revealed a generic benefit for 3′ pairing of increasing lengths, with the median fold-change in relative K_D value in comparison to mismatched seed pairing alone increasing from 1.8- to 3.4- to 36.3-fold as pairing length increased from 4 to 7 to 11 bp, respectively (Figure 2A). Moreover, as 3′-pairing lengths increased, a larger percentage of the more effective 3′ sites exhibited more improved binding affinity than might have been expected based on distributions observed for shorter 3′ sites, indicating preferences for pairing positions and offsets that became more prominent with greater complementarity to the miRNA 3′ end.

To explore these preferences, we identified the pairing position associated with the highest-affinity 3′ site at each length and examined the relative affinities for pairing at that position over a range of pairing offsets (Figure 2B). Nearly all possibilities examined had values readily distinguished from the log-averaged value for seed-mismatched sites alone, with compensatory pairing to miRNA nucleotides 11–16 at optimal offsets yielding binding affinities comparable to that of the canonical 6mer (Figure 2B, left). Further inspection of longer 3′ sites underscored the conclusion that pairing to the GGUUGUA segment spanning positions 11–17 of let-7a is the most consequential for 3′-supplementary pairing, as all optimal pairing positions for 3′ sites ≥7 nt in length paired to this segment. Moreover, inspection of the optimal positions for shorter sites showed that pairing to the 5′ end of this segment (containing the sequence GGUU) was more impactful than pairing to its 3′ end (Figure 2B, right). In addition, increasing the length of pairing from 4 to 11 bp led not only to increased binding affinity at almost all offsets, as might have been expected, but also led to a shift in the optimal offset, with a preferred offset of +2 nt when pairing with only 4 bp and an offset centering on +4 nt when pairing to optimal 3′ sites with 9–11 bp (Figure 2B, left). These length and offset preferences were also observed when examining results of the let-7a replicate experiment (Figures S2A and S2B). Moreover, similar

analyses with the canonical sites yielded no strong positional preferences (Figures S2C–S2E), consistent with the interpretation that the offset preferences observed for the 3′ sites were informed predominantly by binding events that included seed-pairing to the programmed mismatch sites.

To investigate the underpinnings of the change in preferred offset, we plotted the relative affinities of all possible positions, lengths, and offsets for let-7a 3′ pairing (Figure 2C). As pairing length increased beyond 6 bp, two distinct trends emerged: one with a maximal offset of +4 nt and higher-affinity relative K_D values, and another with a maximal offset of +1 nt and more modest relative K_D values. These two offset trends indicated two distinct binding modes. Moreover, the maximal offset of +4 nt nearly always occurred for configurations that included pairing to the G at position 11 of let-7a, with an abrupt switch from the preferred offset of +1 nt to a preferred offset of +4 nt when 3′ pairing began at position 11 rather than 12. These results suggested that pairing to position 11 in the central region of the miRNA is less accessible than pairing to position 12, and therefore a longer loop in the target sequence is required to bridge seed pairing with 3′ pairing that includes position 11. Nonetheless, when pairing to position 11 is enabled through this second binding mode, substantially greater affinity can be achieved.

Some of the lowest relative affinity values were observed for extended 3′-pairing possibilities that began at position 9 with an offset of 0 nt (Figure 2B and C, asterisks). These low values were attributed to AGO2-catalzyed slicing of molecules with extensive contiguous pairing, which depleted these molecules from our bound library. Supporting this idea, analogous sites with offsets of either −1 or +1 nt, which were expected to disrupt slicing due to single-nt bulges in either the miRNA or the site, respectively, did not have aberrantly low relative affinities. Our observation of some slicing during the course of the binding experiment was

consistent with reports that AGO2 can slice sites that have a seed mismatch but are otherwise extensively paired to the guide RNA (Becker et al., 2019; Chen et al., 2017; Wee et al., 2012).

We next used heat maps to visualize the interplay between 3′-site position and pairing length at different offsets (Figure 2D). At the optimal offset length of +4 nt, pairing to let-7a positions 10–20 conferred an ~380-fold increase in affinity over the average seed-mismatched site alone (Figure 2D), leading to an overall binding affinity rivaling that of the canonical 8mer (Figure 2B). The binding affinity of this site and all other sites decreased in a uniform manner with increasing offset values beyond +4 nt. With the exception of pairing configurations beginning at position 12, the binding affinities also uniformly decreased when decreasing the offset value from +4 nt, which further underscored the dominance of the +4-nt binding mode for let-7a.

Previous low-throughput measurements of the benefit of 3′ pairing for let-7a examined the influence of pairing to miRNA positions 13–16 at an offset of 0 nt and found that this pairing conferred a 1.6–2-fold increase in binding affinity (Salomon et al., 2015; Wee et al., 2012). Likewise, our measurements for this 4-nt 3′ site indicated that it conferred a 1.5-fold increase in affinity (Figure 2D). Furthermore, maintaining the offset of 0 nt and the pairing position of 13 and extending pairing to the very 3′ end of let-7a improved the binding affinity to only 3.1-fold (Figure 2D). These results highlight the importance of both the +4-nt offset and pairing to position 11 of let-7a—two features that would have been difficult to identify without comprehensive investigation of the 3′-pairing preferences of this miRNA. Indeed, the importance of these two features is not revealed in an analysis of a dataset that reports the affinities ~23,000 different sites to let-7a because these ~23,000 sites were not designed to analyze the combined effects of varying both pairing position and pairing offset (Becker et al., 2019) (Figure S3).

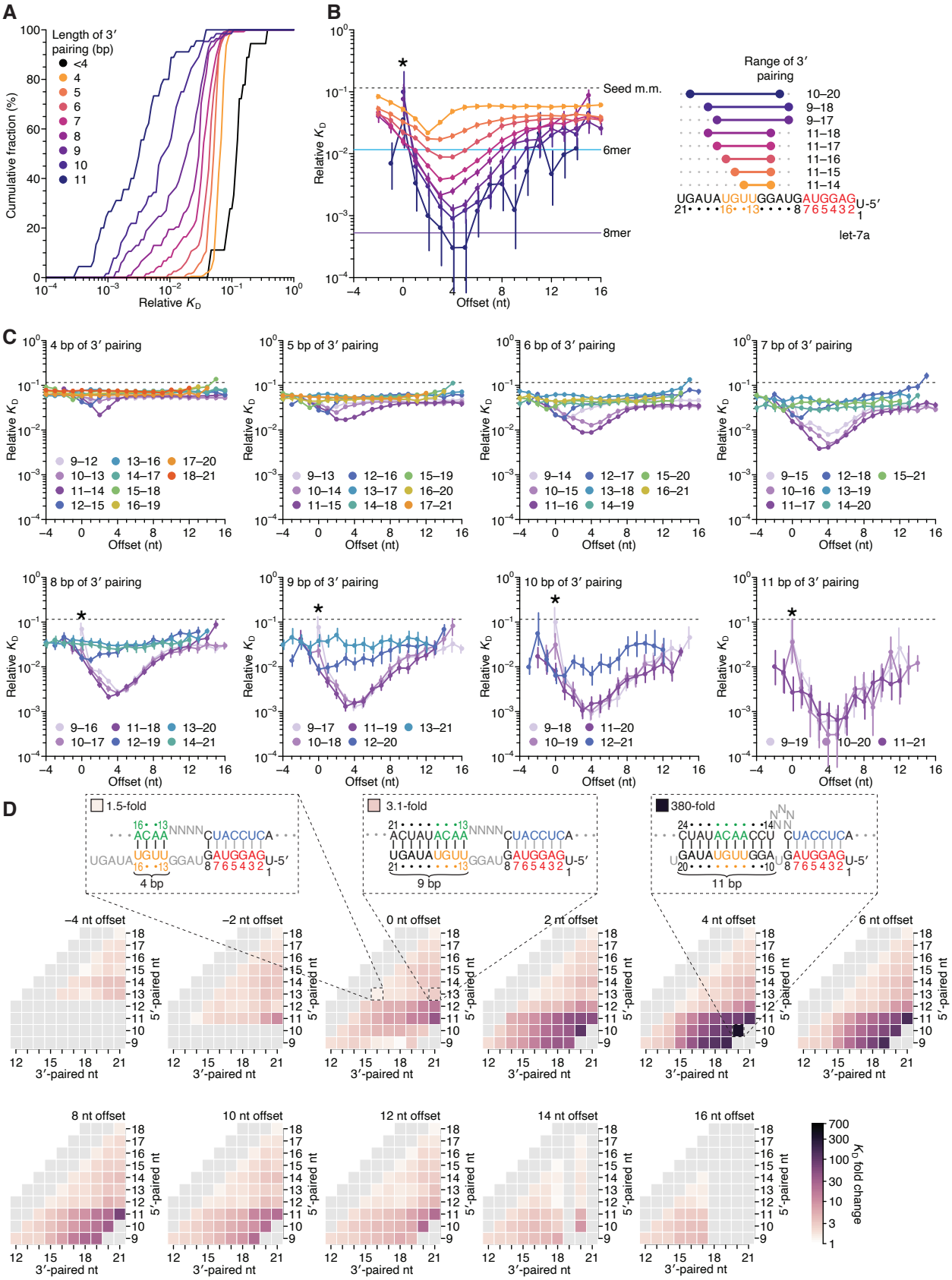

A
Length of 3′ pairing (bp)
<4
4
5
6
7
8
9
10
11
Cumulative fraction (%)
Relative K_D
B
Relative K_D
Seed m.m.
6mer
8mer
Offset (nt)
Range of 3′ pairing
10–20
9–18
9–17
11–18
11–17
11–16
11–15
11–14
UGAUAUGUUGGAUGAUGGAG U-5′
21 16 13 8 7 6 5 4 3 2 1
let-7a
C
4 bp of 3′ pairing
9–12 13–16 17–20
10–13 14–17 18–21
11–14 15–18
12–15 16–19
5 bp of 3′ pairing
9–13 12–16 15–19
10–14 13–17 16–20
11–15 14–18 17–21
6 bp of 3′ pairing
9–14 12–17 15–20
10–15 13–18 16–21
11–16 14–19
7 bp of 3′ pairing
9–15 12–18 15–21
10–16 13–19
11–17 14–20
8 bp of 3′ pairing
9–16 11–18 13–20
10–17 12–19 14–21
9 bp of 3′ pairing
9–17 11–19 13–21
10–18 12–20
10 bp of 3′ pairing
9–18 11–20
10–19 12–21
11 bp of 3′ pairing
9–19 10–20 11–21
D
1.5-fold
ACAA NNNN CUACCUCA
UGAUA UGUU GGAU GAUGGAG U-5′
4 bp
3.1-fold
ACUAUACAA NNNN CUACCUCA
UGAUAUGUU GGAU GAUGGAG U-5′
9 bp
380-fold
CUAUACAACCU CUACCUCA
UGAUAUGUUGGA GAUGGAG U-5′
11 bp
–4 nt offset
–2 nt offset
0 nt offset
2 nt offset
4 nt offset
6 nt offset
8 nt offset
10 nt offset
12 nt offset
14 nt offset
16 nt offset
5′-paired nt
3′-paired nt
K_D fold change
700
300
100
30
10
3
1

Figure 2. Pairing to nucleotide 11 and a +4-nt offset promote high-affinity binding to let-7a. (**A**) Cumulative distributions of relative K_D values for let-7a 3′-compensatory sites that have 4 (orange) to 11 (dark blue) contiguous base pairs of 3′ pairing. Each relative K_D value corresponds to a single length, position, and offset of 3′-compensatory pairing, and was calculated after aggregating the read counts of all 18 possible seed-mismatch types at the programmed region of the library. For comparison, the distribution for sites with <4 bp of contiguous 3′ pairing is also shown (black); for this distribution relative K_D values of each of the 18 seed-mismatch types were calculated separately. (**B**) Relative K_D values of let-7a 3′-compensatory sites that had optimally positioned 3′ pairing of lengths 4–11 bp. For each of these 3′-pairing lengths, the position associated with the greatest affinity is shown (right), and the relative K_D values of the 3′-compensatory sites at each measured offset are plotted (left). Vertical lines indicate 95% confidence intervals. The dashed horizontal line indicates the geometric mean of the 18 relative K_D values of the seed mismatch sites, each calculated from reads with <4 nt of contiguous complementarity to the miRNA 3′ region. The horizontal blue and purple lines indicate the relative K_D values of the canonical 6mer and 8mer sites, respectively. The asterisk denotes anomalously low binding affinity observed for pairing at position 9 with an offset of 0 nt. (**C**) The dependency of let-7a 3′-pairing affinity on pairing length, position, and offset. Each panel shows the relative K_D values for 3′ pairing of a specified length over a range of positions and offsets. Each trend line is colored according to pairing position, spanning positions 9 (light violet) to 18 (red) when possible. Otherwise, these panels are as in (B, left). (**D**) Affinity profile of the let-7a 3′ region. Each cell indicates the fold-change in relative K_D attributed to a 3′ site with indicated length, position, and offset of pairing. Each row within a heat map corresponds to a different miRNA nucleotide at the start of the 3′ pairing, and each column corresponds to a different miRNA nucleotide at the end of the 3′ pairing. Each heat map shows the results for a different offset. The three diagrams indicate the fold-change values and architectures for 3′ sites pairing to miRNA nucleotides 13–16 with an offset of 0 nt (left), pairing to miRNA nucleotides 13–21 with an offset of 0 nt (middle), and pairing to miRNA nucleotides 10–20 with an offset of +4 nt (right). Gray boxes indicate pairing ranges that were either too short (<4 bp) or too long (>11 bp) for relative K_D values to be reliably calculated. Black vertical lines depict perfect Watson–Crick pairing, and gray vertical lines indicate a non-Watson–Crick match somewhere within the seed pairing.

Different miRNAs have distinct 3′-pairing preferences

The optimal 3′-pairing architecture for let-7a differed from that previously elucidated for miRNAs more generally (Grimson et al., 2007). When pooling repression and conservation data for 11 miRNAs, pairing to miRNA nucleotides 13–16, with an offset of 0 nt appears to be most consequential (Figure 1A) (Grimson et al., 2007). Because the previous analysis represents the average of trends derived from multiple miRNAs, a diversity of miRNA-specific 3′-pairing preferences, analogous to the observed diversity of seed-pairing preferences (McGeary et al.,

2019), might explain this disagreement. We therefore measured the 3′-pairing profiles of two other miRNAs, miR-1 and miR-155, for comparison to the let-7a profile. As with let-7a, we synthesized programmed libraries enriched for all possible single-nucleotide seed mismatches at positions 2–7, performed AGO-RBNS, and calculated relative K_D values for 3′ sites 4–11 nt in length, at all possible positions and offsets present within the library.

Stabilizing 3′ pairing was observed for both miR-1 (Figures 3A and 3B) and miR-155 (Figures 3C and 3D), with binding affinity increasing with the length of pairing, as observed for let-7a (Figure 2). However, the magnitude of increased binding affinity differed from that of let-7a and that of each other: the affinity of 3′ pairing to miR-1 was more modest, with only a handful of 11-bp pairing possibilities reaching affinity comparable to that of the canonical 6mer site (Figure 3A), whereas for miR-155, most 8-bp pairing possibilities achieved such affinity (Figure 3C). The positions of the best sites at each length also differed from let-7a. For miR-1, optimal 4-bp sites paired to miRNA nucleotides 12–15, and as optimal sites increased in length, pairing extended continuously, primarily towards the 3′ end of the miRNA and never reaching to miRNA nucleotide 10 (Figure 3B, right). By contrast, for miR-155, optimal 4-bp sites paired to miRNA nucleotides 13–16, and as optimal sites increased in length, pairing sometimes shifted discontinuously and never included miRNA nucleotide 12 (Figure 3D, right).

Analysis of each of the optimal 3′ sites of miR-1 and miR-155 along the length of the random region indicated that, unlike sites for let-7a, those for neither of these two miRNAs underwent a significant shift in the preferred offset (Figures 3B and 3D, left). Nevertheless, the offset preferences of miR-1 did become more tolerant of a wider range of positive values, consistent with a minor contribution of an alternative binding mode resembling that of let-7a. The offset preferences of miR-155 substantially diminished with increased pairing. These reduced offset preferences coincided with pairing to the $G_{19}G_{20}G_{21}G_{22}$ stretch near the 3′ end of

miR-155 and might relate to the ability of this miRNA to participate in seed-autonomous 3′-pairing, as detected when performing AGO-RBNS with fully randomized RNA libraries (McGeary et al., 2019). However, distinguishing between seed-autonomous pairing and seed-dependent, offset-agnostic 3′-compensatory pairing was not possible using our results, due to the presence of a seed-mismatched site in each molecule of the programmed library.

In summary, the most optimal 3′ sites each paired to at least two nucleotides of the miRNA segment spanning positions 13–16, which was previously identified as most consequential for 3′ pairing, but frequently did not pair to the entire segment. Shorter optimal sites consistently preferred pairing to G nucleotides adjacent to miRNA nucleotides 13–16. For example, shorter optimal sites to let-7a paired to the $G_{11}G_{12}$ sequence element 5′ of this segment rather than to positions 15 and 16 (Figure 2B, right), the optimal 4-nt site to miR-1 paired to G_{12} rather than to position 16 (Figure 3B, right), and intermediate-length optimal sites to miR-155 paired to $G_{19}G_{20}G_{21}G_{22}$ rather than to positions 13 and 14 (Figure 3D, right). These trends were also observed when comprehensively examining all possible positions, lengths, and offsets for miR-1 and miR-155 (Figure S4). In aggregate, these results supported the report of an intrinsic preference for pairing to miRNA nucleotides 13–16 (Grimson et al., 2007) but also indicated that the miRNA sequence imparts additional preferences, resulting in unanticipated differences between the optimal sites of individual miRNAs. These sequence-specific preferences tended to favor pairing to G residues of the miRNA, which was presumably explained by the greater stability of G:C pairing over A:U pairing, although the presence of only a single C nucleotide prevented investigation of a primary-sequence preference among the 3′ regions of these three miRNAs. We also observed differences between miRNAs in the strength of 3′ pairing. Compared to 3′-site affinities observed for let-7a, affinities were substantially lower for miR-1 and substantially greater for miR-155 (median K_D fold-change values with 11 bp of 3′ pairing,

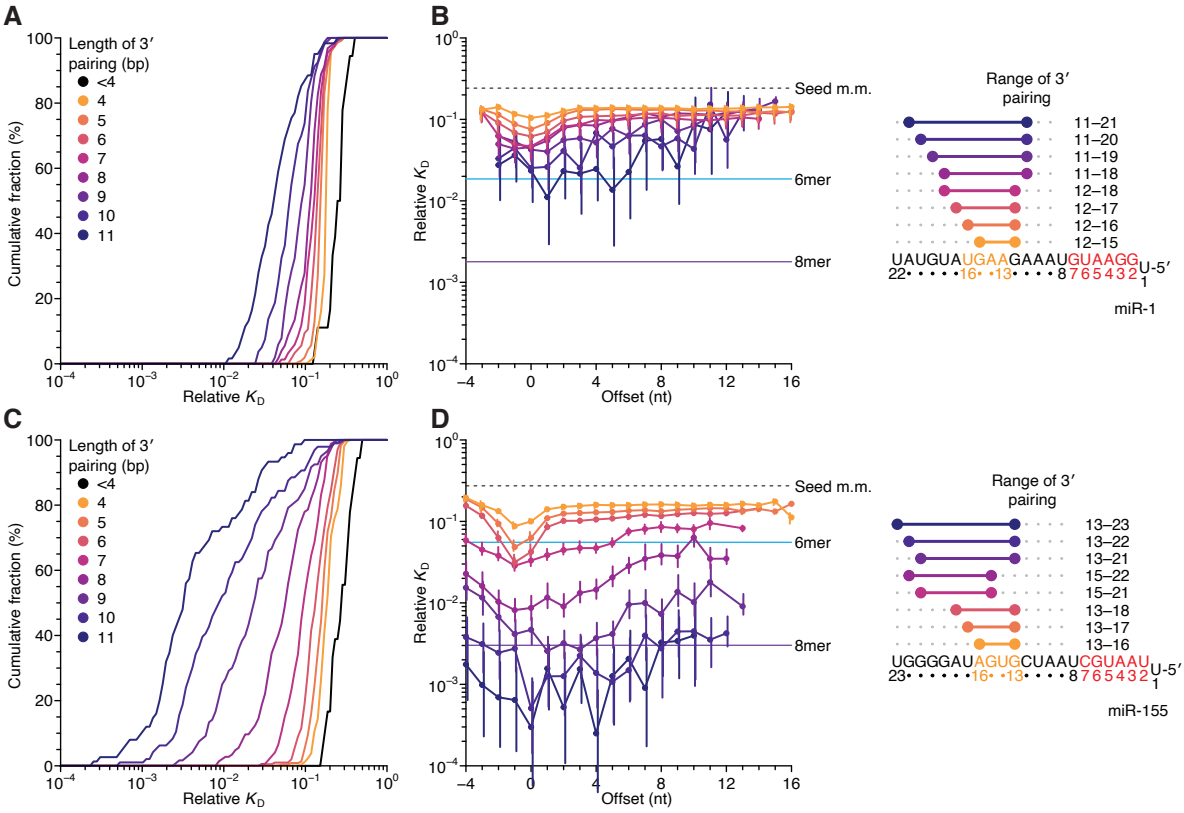


Figure 3. Relative affinity measurements of 3′-compensatory sites of miR-1 and miR-155. (**A**) Cumulative distributions of relative K_D values for miR-1 3′-compensatory sites. Otherwise, this panel is as in Figure 2A. (**B**) Relative K_D values of miR-1 3′-compensatory sites that had optimally positioned 3′ pairing of lengths 4–11 bp. Otherwise, this panel is as in Figure 2B. (**C**) Cumulative distributions of relative K_D values for miR-155 3′-compensatory sites. Otherwise, this panel is as in Figure 2A. (**D**) Relative K_D values of miR-155 3′-compensatory sites that had optimally positioned 3′ pairing of lengths 4–11 bp. Otherwise, this panel is as in Figure 2B.

36, 5.8, and 133 for let-7a, miR-1 and miR-155, respectively). Thus, our results indicated that the loading of the guide RNA into the AGO protein does not fully standardize either the architecture of optimal 3′ pairing or the magnitude of its benefit.

Pairing and offset coefficients describe unique 3′-pairing profiles for each miRNA

To summarize the results for miR-1 and miR-155, we generated heat maps representing the binding affinity at all possible pairing positions for all pairing lengths of 4–11 bp, as a function of pairing offset (Figure S5), as with let-7a (Figure 2C). Within each heat map, adjacent cells

corresponded to the difference in K_D fold change caused by the addition or removal of a pair at either the 5′ end (adjacent rows) or the 3′ end (adjacent columns) of the 3′ site, while maintaining the same offset. The similarities observed between heat maps for the same miRNA at different offsets indicated that each change in offset altered the binding affinity of all 3′-pairing possibilities in a consistent manner, which in turn indicated that for each of the three miRNAs, the effect of pairing offset was largely independent of the effect of guide–target complementarity (Figures 2C and S5).

To test this independence, we examined the extent to which the affinities could be quantitatively explained as a simple function that considered the contribution of the pairing range, which was defined by pairing position and length, as modified by the contribution of the pairing offset. Our model explained the data well (r^2 = 0.92, 0.86, and 0.96 for let-7a, miR-1, and miR-155, respectively, Figure S6), and yielded a set of pairing and offset coefficients for each miRNA. Each pairing coefficient represented the ΔG of the corresponding pairing range at its optimal offset, and each offset coefficient represented the reduction in ΔG observed at suboptimal pairing offsets (Figures 4A–4C). For each miRNA, the pairing coefficients corresponded well with the affinities observed at the preferred offset (Figures 4A–4C, r^2 = 0.98, 0.97, and 0.96, respectively). Moreover, these coefficients, which reported on the ensemble behavior observed over all 934, 1061, and 1180 K_D values measured for let-7a, miR-1, and miR-155, respectively, quantitatively captured the qualitative observations made earlier from analysis of subsets of the data. For example, they captured the respective importance of pairing to nucleotides 11, 12, and 20 and the respective preferences for offsets of +4, +1, and +1 nt for let-7a, miR-1, and miR-155, respectively. They also captured the more narrowed offset preferences of let-7a in comparison to those of miR-1 and miR-155 (Figures 4A–4C, middle-left) and the contribution of pairing starting at miRNA position 15 for miR-155 (Figure 4C, left). Moreover,

the high agreement of the pairing and offset coefficients of let-7a with those determined independently from the let-7a replicate experiment (r^2 = 0.994 and 0.988, respectively; data not shown) indicated that these coefficients were determined with minimal experimental error.

Because the pairing coefficients represented the thermodynamic benefit of each pairing possibility, we examined how well each set of pairing coefficients was explained by nearest-neighbor rules that predict the stability of RNA hybridization in solution. To do so, we calculated the predicted ΔG value for each 3′ site pairing to the miRNA 3′ region (Figure 4D) and adjusted each value by subtracting the mean value for that length of pairing, which was done to remove the trivial effect of increasing pairing length (Figure 4E). When comparing these length-adjusted values with analogously adjusted pairing coefficients, we observed a strong relationship for both let-7a and miR-155, which explained most of the variation in the length-adjusted coefficients, and a much weaker relationship for miR-1. Nevertheless, even when focusing on results for let-7a and miR-155, the apparent effect size was less than that expected by the relationship $\Delta G = -RT \ln K$ (Figure 4E, dashed lines). Thus, as observed with the miRNA seed region (McGeary et al., 2019; Salomon et al., 2015), compared to RNA free in solution, association with AGO reduces the differences in binding energy observed when hybridizing to different miRNA 3′-end sequences.

This reduction in magnitude also applied to the overall contribution of 3′ pairing (Figure S7A). For instance, although the >200-fold differences in binding affinity imparted by the top 11-nt 3′ sites of let-7a and miR-155 might seem large, the ΔG predicted for each of these sites was −14.8 kcal/mol and −20.1 kcal/mol, which corresponded to respective fold differences of 2.7×10^{10} and 1.5×10^{14}. Presumably the benefit of pairing to 3′ sites was mostly offset by the cost of disrupting favorable interactions between unpaired 3′ regions and AGO, as has also been proposed in the context of siRNA-mediated target cleavage (Tomari and Zamore, 2005). The

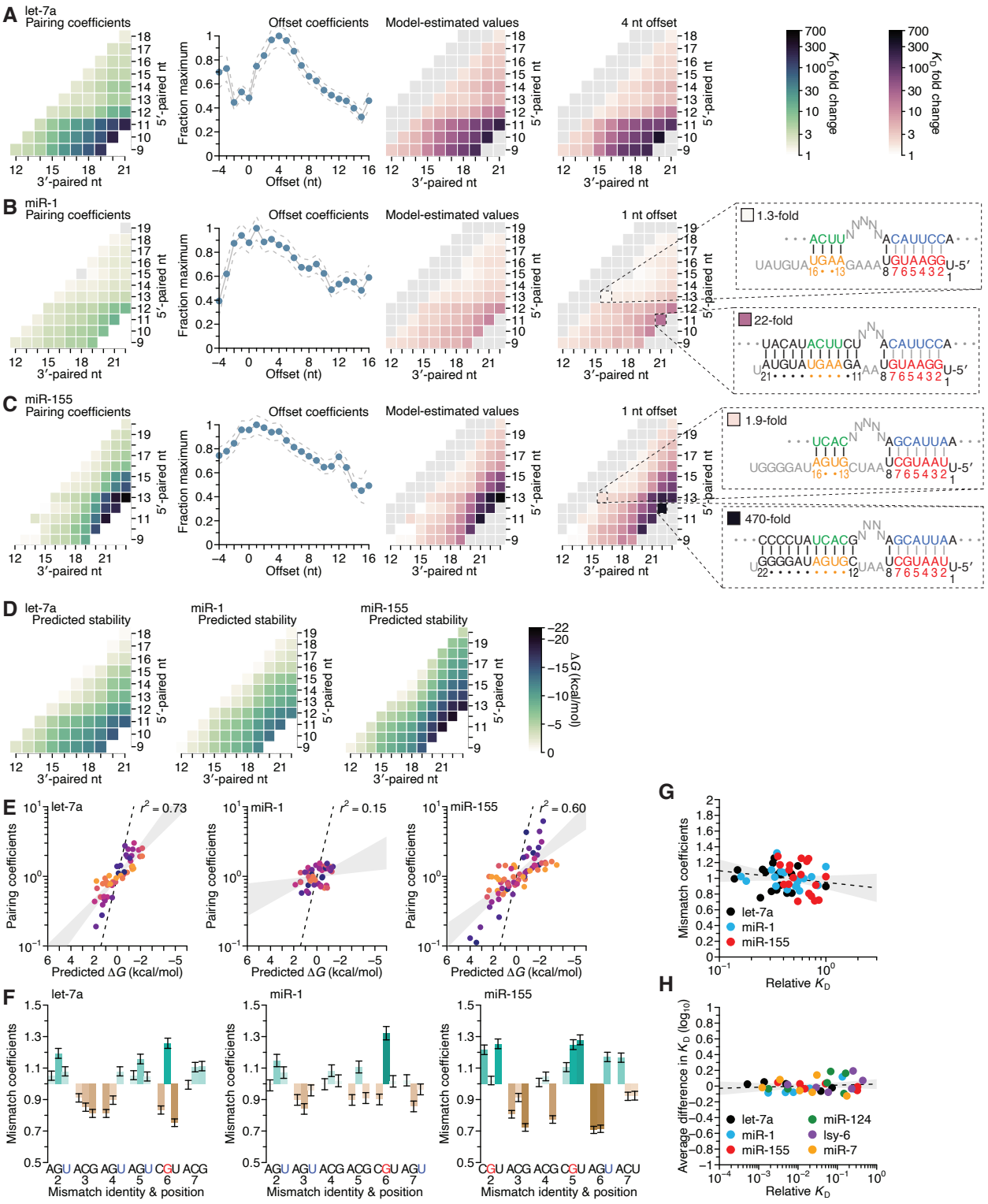


Figure 4. Distinct pairing-range, offset, and seed-mismatch preferences of different miRNAs.

(**A**–**C**) Model-based analyses of 3′-pairing preferences of let-7a (A), miR-1 (B), and miR-155 (C). For each miRNA, 3′-pairing affinities are described by a set of pairing coefficients (left) and offset coefficients (middle, left; dashed lines, 95% confidence interval), which when multiplied

together (middle, right) recapitulated measured K_D fold-change values (right, let-7a values replotted from Figure 2C). The parameters were obtained by maximum-likelihood estimation with a nonlinear energy model. For both miR-1 (B) and miR-155 (C), the two pairing diagrams indicate the fold-change value and architecture for a 3′ site pairing to miRNA nucleotides 13–16 (top) in comparison to the fold-change value and architecture of the 3′ site with the greatest measured affinity (bottom), both at the optimal offset of +1 nt. Pairing coefficients, model predictions, and K_D fold-change values of miR-1 were not calculated for pairing to miRNA positions 15–18 and 19–22 because these two segments were identical (gray boxes). (**D**) Predicted ΔG values of the 3′ sites with pairing coefficients in (A–C). (**E**) The relationship between the model-derived pairing coefficients (A–C) and the predicted ΔG values (D). Points are colored according to pairing length, as in Figure 2A. To control for the trivial effect of increasing pairing length, pairing-range coefficients were divided by the geometric mean of all coefficients with the same length, and ΔG values of each length were normalized to the mean ΔG value of pairings with the same length. The gray region represents the 95% confidence interval of the relationship when fitting a linear model to the data (r^2, coefficient of determination), and the dashed line represents the predicted thermodynamic relationship given by $K = e^{-G/RT}$. (**F**) Distinct effects of seed mismatches on 3′-pairing affinities of let-7a, miR-1, and miR-155. For each miRNA, seed-mismatch coefficients were derived by maximum-likelihood estimation, fitting a nonlinear model to the K_D fold-change values observed when examining 3′-site enrichment separately for each of the 18 seed mismatches. The error bars indicate 95% confidence intervals. Wobble pairing in which the G was in either the miRNA or the target is indicated in blue and red, respectively. (**G**) Relationship between affinity of 3′-compensatory pairing and that of seed-site binding. For each seed mismatch, the coefficient from (F) is plotted as a function of the relative K_D value of that mismatch, as measured using results from the programmed libraries for let-7a (black), miR-1 (blue), and miR-155 (red). The dashed line shows the linear least-squares fit to the data, with the gray interval indicating the 95% confidence interval. (**H**) Relationship between affinity of 3′-supplementary pairing and that of seed-site binding. For each of the six seed-matched site types (Figure 1A, left) and for each of the six miRNAs (key), the relative affinity of the top quartile of all 4- and 5-nt 3′ sites with their preferred offsets is plotted as a function of the relative affinity of the seed-matched site. Relative affinities were measured from analysis of previous AGO-RBNS that used a random-sequence library (Figure S13).

magnitude of this inferred cost appeared specific to each miRNA, implying that AGO might have some sequence preferences when interacting with unpaired miRNA 3′ regions. For example, pairing to either nucleotides 9–19 of let-7a or nucleotides 11–21 of miR-1 was predicted to occur with equivalent ΔG values of −13.5 kcal/mol, yet the model-determined contributions of these sites were 160- and 14-fold, respectively (Figure S7A, left and middle).

Separating the comparison between K_D fold-change and ΔG based on whether the contiguous range of pairing included nucleotide 11, 12, and 20 for let-7a, miR-1, and miR-155, respectively, revealed a nonlinear benefit of pairing to these nucleotides (Figures S7B and S7C), such that their inclusion within the 3′ pairing enabled the other paired nucleotides to contribute more to the interaction. We also note that using the measured affinities rather than pairing coefficients did not increase agreement with ΔG (Figures S7D and S7E), suggesting that the use of the pairing coefficients did not lead to loss of information contained within the data from which they were generated.

The success of our analyses of data obtained from programmed libraries prompted analysis of data obtained previously from fully randomized libraries (McGeary et al., 2019) (Figures S8A–S8E). For let-7a, miR-1, and miR-155, the pairing and offset coefficients derived from data from the two types of libraries agreed well with each other, provided that 3′-pairing lengths did not extend beyond 8 bp (Figures S8G and S8H). However, when pairing lengths extended beyond 8 bp, affinity values were not reliably determined because the sites were only sparsely represented in the input libraries. Inspection of pairing preferences of these three miRNAs, as indicated by their pairing and offset coefficients derived from the random-library data, revealed their distinguishing features, including: the importance of pairing to position 11 of let-7 and position 12 of miR-1, the right-shifted preferred offset of let-7, and the relative ordering of the maximal benefit of 3′ pairing, with that of miR-155 exceeding that of let-7a, which exceeded that of miR-1 (Figures S8A–S8C).

Having determined the utility and limits of analyses of data from fully randomized libraries, we turned to the analyses of 3′ pairing to miR-124, lsy-6, and miR-7, for which data from programmed libraries was not available. These analyses showed that miR-124, like let-7a, had both preferred pairing to position 11 and a right-shifted preferred offset of pairing (Figure

S8D). To look for evidence of multiple binding modes within these original AGO-RBNS datasets, we repeated the analyses of both Figure 2B (for pairing lengths of 4–8 bp) and Figure 2C (for pairing lengths of 4 and 5 bp), using the original AGO-RBNS data for miR-124, lsy-6, and miR-7 (Figure S9). For comparison, we also repeated these analyses using the original AGO-RBNS data for let-7a, for which we had evidence of two binding modes from the programmed-library AGO-RBNS data. For each of the four miRNAs, we found evidence of two binding modes. Both let-7a and miR-124 had the previously observed pattern, in which the binding mode with the positive offset and paring to nucleotide 11 had binding affinity greater than that of the binding mode with an offset of 0 nt and pairing to only nucleotide 12 (Figures S9A–S9D). However, lsy-6 and miR-7 had a different pattern, in which the binding mode corresponding to the positive offset and pairing to nucleotide 11 had binding affinity similar to that of the binding mode with an offset of 0 nt and pairing to only nucleotide 12 (Figures S9E–S9H).

These examples provided further evidence of a second binding mode, in which productive 3′ pairing extended to nucleotide 11, provided that additional unpaired target nucleotides were available to bridge pairing between the seed and this nucleotide. These results also suggested that pairing to the $G_{11}G_{12}$ dinucleotide found in both the let-7a and miR-124 sequences enabled this second binding mode to dominate over the first, whereas pairing to the single G_{11} found in lsy-6 and miR-7 added to site affinity but did not enable the second binding mode to dominate. Indeed, although miR-7 appeared to have both binding modes, it had the weakest 3′-compensatory pairing of the six miRNAs profiled, with 8-nt 3′ sites never contributing more than an 18-fold increase in binding affinity.

The analyses of the miR-124 and lsy-6, which each had multiple C nucleotides in their 3′ region, allowed us to return to the question of whether pairing to miRNA G nucleotides might be favored over pairing to C nucleotides. Pairing to C_{15} of lsy-6 substantially added to binding

affinity. For example, the 4.2-fold greater affinity of the position-12–15 site over the position-11–14 site indicated that pairing to C_{15} was favored over paring to G_{11}, and extending pairing from positions 11–14 to 11–15 increased affinity 8.2-fold (Figure S8E). Pairing to C_{13} was also somewhat preferred, as illustrated by the 1.8-fold greater affinity of the position-13–17 site over the position-14–18 site, and the 3.2-fold benefit of extending pairing from positions 14–18 to 13–18. However, pairing to $C_{19}C_{20}$ of miR-124 did not seem to have the same impact as pairing to $G_{19}G_{20}$ of miR-155, as illustrated by the negligible (0.9-fold) benefit of extending the miR-124 pairing from positions 13–18 to 13–20, compared to the 14-fold benefit for miR-155. These results supported the idea that pairing to a G in the miRNA 3′ region is generally favored over pairing to a C, although pairing to a C centrally located within the 3′ region can be impactful.

The type of seed mismatch affects the affinity of 3′ pairing

To examine the influence of seed-mismatch position and identity, we analyzed the full set of 16,235, 18,076, and 19,666 K_D values of let-7a, miR-1, and miR-155, no longer combining read counts for the 18 possible seed-mismatch sites in the programmed library prior to K_D estimation. For each pairing, offset, and seed-mismatch possibility, the relative K_D value of the 3′-compensatory site was divided by that of its seed-mismatch site to generate a fold-change value representing the contribution of the 3′ site to affinity. An expanded model was then fit to these data, in which the log(K_D fold change) was described as the product of its pairing, offset, and seed-mismatch coefficients. The seed-mismatch coefficients were modeled to influence the affinity of 3′ pairing as a function of the amount of 3′-pairing affinity that was attainable, which varied between miRNAs. Thus the range of 0.50, 0.48, and 0.57 observed for seed-mismatch coefficients for let-7a, miR-1, and miR-155, respectively (Figure 4F), corresponded to 9.2-, 2.6-, and 11.2-fold predicted variation in binding affinity for each of the respective miRNAs in the

context of its most favorable pairing and offset, with the lower impact on miR-1 attributed to the lower amount of affinity attainable by its 3′ pairing. These predicted effects generally agreed with those observed when examining affinities of the top quartile of 3′ sites in the context of their optimal offset, which respectively varied by 14.9-, 4.6-, and 6.5-fold depending on seed-mismatch identity. Furthermore, visual inspection of the trends in observed 3′-site affinities confirmed the increased effect of seed mismatches for higher-affinity 3′ sites (Figures S10–S12). For example, only a few of the 4-nt 3′ sites to let-7a were sensitive to the particular seed-mismatch type (Figure S10A), whereas for 8-nt sites, more positions and offsets exhibited such variation, and these were positions and offsets with higher average affinities (Figure S10E).

The affinity of seed-mismatch sites lacking 3′ pairing had little relationship with the influence of the mismatch on 3′-pairing affinity (Figure 4G). Likewise, examination of data from the six random-library AGO-RBNS experiments found no relationship between the canonical site affinities of sites lacking 3′ pairing and the influence of the site on the binding contributed by the top 4- and 5-nt 3′ sites (Figures 4H and S13). Furthermore, the average effect of canonical site type on 3′ binding affinity was small, with only six out of the 36 miRNA–site combinations having a >0.1 effect on $\log_{10}(K_D$ fold change), corresponding to an ~25% change in binding affinity. Together, these results indicate that for 3′-supplementary pairing, the benefit of the 3′ pairing is largely the same between sites, but that for 3′-compensatory pairing, the potential benefit of 3′ pairing is differentially available depending on the identity of the seed mismatch. This might be due to a differential ability of these mismatches to elicit a conformational change in AGO allowing pairing to the 3′ end (Schirle et al., 2014; Sheu-Gruttadauria et al., 2019a). Alternatively, some sites may have dwell times shorter than that required to establish pairing to the miRNA 3′ region.

When comparing the effects for guide–target nucleotide possibilities, strong trends did not emerge within miRNAs (e.g., when comparing the effects of a mismatched A, G, or U to the G at position 2 with those of the mismatches to the G at position 4 of let-7a), or between miRNAs (e.g., when comparing of effects of mismatches to the G at position 3 of miR-1 with those to the G at position 6 of miR-155). However, in cases in which the same nucleotide occurred at the same position for two different miRNAs, some correspondence was observed (positions 2 and 6 of let-7a and miR-1, position 3 of let-7a and miR-155, position 4 of miR-1 and miR-155). Notably, the miRNA–target U:G mismatch at position 6, which was the most favored mismatch for both let-7 and miR-1, occurs within one of the two compensatory sites within the 3′ UTR of *C. elegans lin-41*, consistent with the idea that the mismatch effects observed by RBNS are of consequence for cellular targeting.

The seed-mismatch and 3′-sequence effects act independently

The distinct pairing, offset, and seed-mismatch preferences of the three miRNAs measured using the programmed libraries raised the question of the extent to which these preferences depended on the sequence of the seed region, the sequence of the 3′ region (i.e., beginning at miRNA nucleotide 9), or a combination of the two. To answer this question, we generated two chimeric miRNAs, one fusing the seed of miR-155 to the 3′ region of let-7a (miR-155–let-7a) and the other fusing seed of let-7a to the 3′ region of miR-155 (let-7a–miR-155) (Figure 5A), and then performed AGO-RBNS using their corresponding seed-mismatched programmed libraries. As done previously with the natural miRNAs, we first determined the pairing and offset preferences of both chimeric miRNAs by summing over all 18 seed mismatch types, measuring the K_D fold change for each range of pairing and offsets possible in the libraries, and fitting a multiplicative

model of pairing and offset-preferences to the resultant 1181 and 934 measured affinity values for let-7a–miR-155 (Figure 5B) and miR-155–let-7a (Figure 5C).

Both of the chimeric miRNAs had 3′-pairing and offset preferences that were remarkably similar to those of the natural miRNAs containing the same 3′ sequences (Figures 4A, 4C, 5B, and 5C). Indeed, comparison of length-normalized pairing and offset coefficients for each chimeric miRNA to those of either its 3′-native or seed-native miRNAs revealed a high correspondence for all four 3′-native comparisons (Figures 5D and 5E) and much lower correspondence for all four seed-native comparisons (Figures 5F and 5G), Furthermore, the fitted slopes for four 3′-native comparisons approached unity (range 0.80–1.17), which showed that the effect sizes of these preferences were similar regardless of whether the coefficients were derived from chimeric or native miRNA datasets.

When analyzing the effects of the 18 seed mismatches on the affinity of 3′ pairing, miRNAs with the same seed sequence but different 3′ sequences had largely similar preferences (Figures 4F, 5H, and 5I), with the most striking differences being the increased affinity in the context of a mismatched A at position 7 of the let-7–miR-155 chimeric miRNA, and decreased affinity in the context of a mismatched U at position 6 for of the miR-155–let-7a chimeric miRNA. Despite these outliers, the influence of the seed mismatch on the magnitude of 3′-pairing affinity depended primarily on the seed-mismatch type and position, with relatively little dependence on the sequence of the 3′ region.

Sequence preferences for 3′ sites are maintained at adjacent positions

We next sought to investigate the positional dependence of the preferences for pairing to particular nucleotides of the 3′ end. To do so, we repeated the AGO-RBNS procedure with let-7a variants that had single-nucleotide insertions and deletions that shifted the let-7a 3′ sequence by a

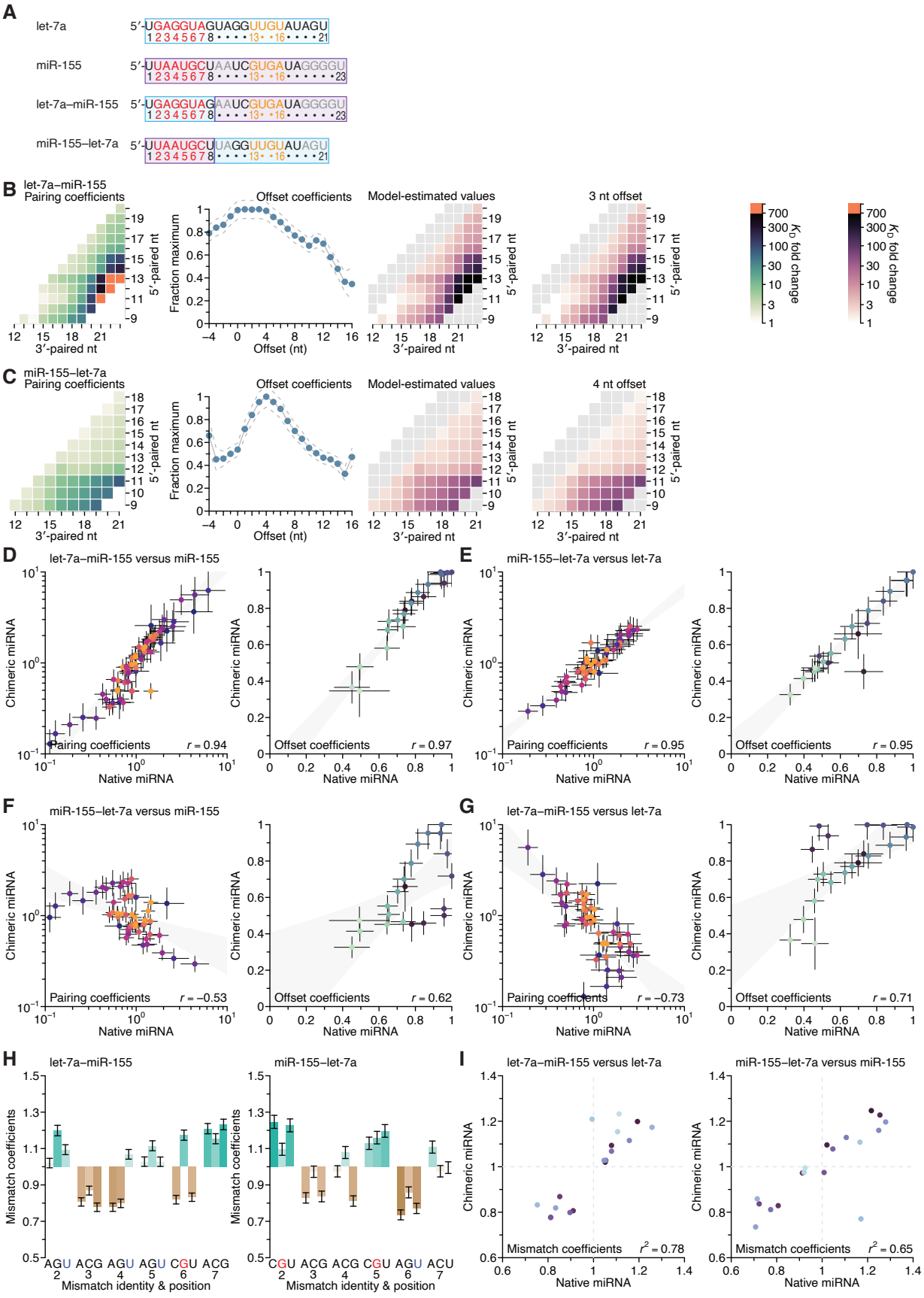

A
let-7a
5′-UGAGGUAGUAGGUUGUAUAGU
miR-155
5′-UUAAUGCUAAUCGUGAUAGGGGU
let-7a–miR-155
5′-UGAGGUAGAAUCGUGAUAGGGGU
miR-155–let-7a
5′-UUAAUGCUUAGGUUGUAUAGU
B
let-7a–miR-155
Pairing coefficients
Offset coefficients
Model-estimated values
3 nt offset
5′-paired nt
3′-paired nt
Fraction maximum
Offset (nt)
K_D fold change
C
miR-155–let-7a
Pairing coefficients
Offset coefficients
Model-estimated values
4 nt offset
D
let-7a–miR-155 versus miR-155
Pairing coefficients r = 0.94
Offset coefficients r = 0.97
E
miR-155–let-7a versus let-7a
Pairing coefficients r = 0.95
Offset coefficients r = 0.95
F
miR-155–let-7a versus miR-155
Pairing coefficients r = –0.53
Offset coefficients r = 0.62
G
let-7a–miR-155 versus let-7a
Pairing coefficients r = –0.73
Offset coefficients r = 0.71
Chimeric miRNA
Native miRNA
H
let-7a–miR-155
miR-155–let-7a
Mismatch coefficients
AGU ACG AGU AGU CGU ACG
CGU ACG ACG CGU AGU ACU
Mismatch identity & position
I
let-7a–miR-155 versus let-7a
Mismatch coefficients r^2 = 0.78
miR-155–let-7a versus miR-155
Mismatch coefficients r^2 = 0.65

Figure 5. Independence of seed-mismatch and 3′-sequence effects.
(**A**) Sequences of native let-7a, native miR-155, a chimeric miRNA containing the seed region of let-7a appended to nucleotides 9–23 of miR-155 (let-7a–miR-155), and a chimeric miRNA containing the seed region of miR-155 appended to nucleotides 9–21 of let-7a (miR-155–let-7a). (**B** and **C**) Pairing and offset coefficients describing the 3′-pairing preferences of let-7a–miR-155 (B) and miR-155–let-7a (C). Orange cells indicate pairing coefficients or K_D fold-change values between 700–1200. Otherwise, this panel is as in Figure 4A. (**D**) Comparison of the pairing and offset coefficients determined for let-7a–miR-155 with those of miR-155. Left, each pairing coefficient was divided by the geometric mean of all pairing coefficients of the same length for that miRNA. Points are colored according to pairing length, as in Figure 2A; error bars indicate 95% confidence intervals. Right, the offset coefficients are colored from light blue to dark blue, progressing from offsets of −4 to +16 nt; error bars indicate 95% confidence intervals. For each graph, the gray region indicates the 95% confidence interval for the linear least-squares fit to the data (*r*, Pearson correlation coefficient). (**E**) Comparison of the pairing and offset coefficients determined for miR-155–let-7a with those of let-7a. Otherwise, this panel is as in (D). (**F**) Comparison of the pairing and offset coefficients determined for let-7a–miR-155 with those of let-7a. Otherwise, this panel is as in (D). (**G**) Comparison of the pairing and offset coefficients determined for miR-155–let-7a with those of miR-155. Otherwise, this panel is as in (D). (**H**) Seed-mismatch coefficients of the let-7a–miR155 (left) and miR-155–let-7a (right) chimeric miRNAs. Otherwise, this panel is as in Figure 4F. (**I**) Correspondence between mismatch coefficients of chimeric miRNAs and those of their seed-native miRNAs. For let-7a–miR-155 (left) and miR-155–let-7a (right), the values from (H) are plotted against those of Figure 4F (r^2, coefficient of determination).

single nucleotide in either direction [let-7a(−1) and let-7a(+1)] while maintaining the miRNA length (Figure 6A). Comparison of the pairing preferences of let-7a(−1) and let-7a(+1) to those of native let-7a indicated that the characteristic benefit of pairing to the G found at nucleotide 11 of the native miRNA was maintained in both variants. Thus, the most consequential nucleotide shifted to 10 when this G shifted to position 10 in let-7a(−1), and likewise, it shifted to 12 for let-7a(+1) (Figures 6B–6D). Pairwise comparison of each of the 36 and 28 pairing possibilities between 4–11 nt shared between let-7a and let-7a(−1) and let-7a(+1), respectively, revealed movement of the consequential nucleotide further 5′ within the miRNA sequence partially reduced the binding affinity, whereas moving it further 3′ had no appreciable effect (Figure 6E). These results suggest that miRNA position 10 might be less accessible than positions 11 or 12.

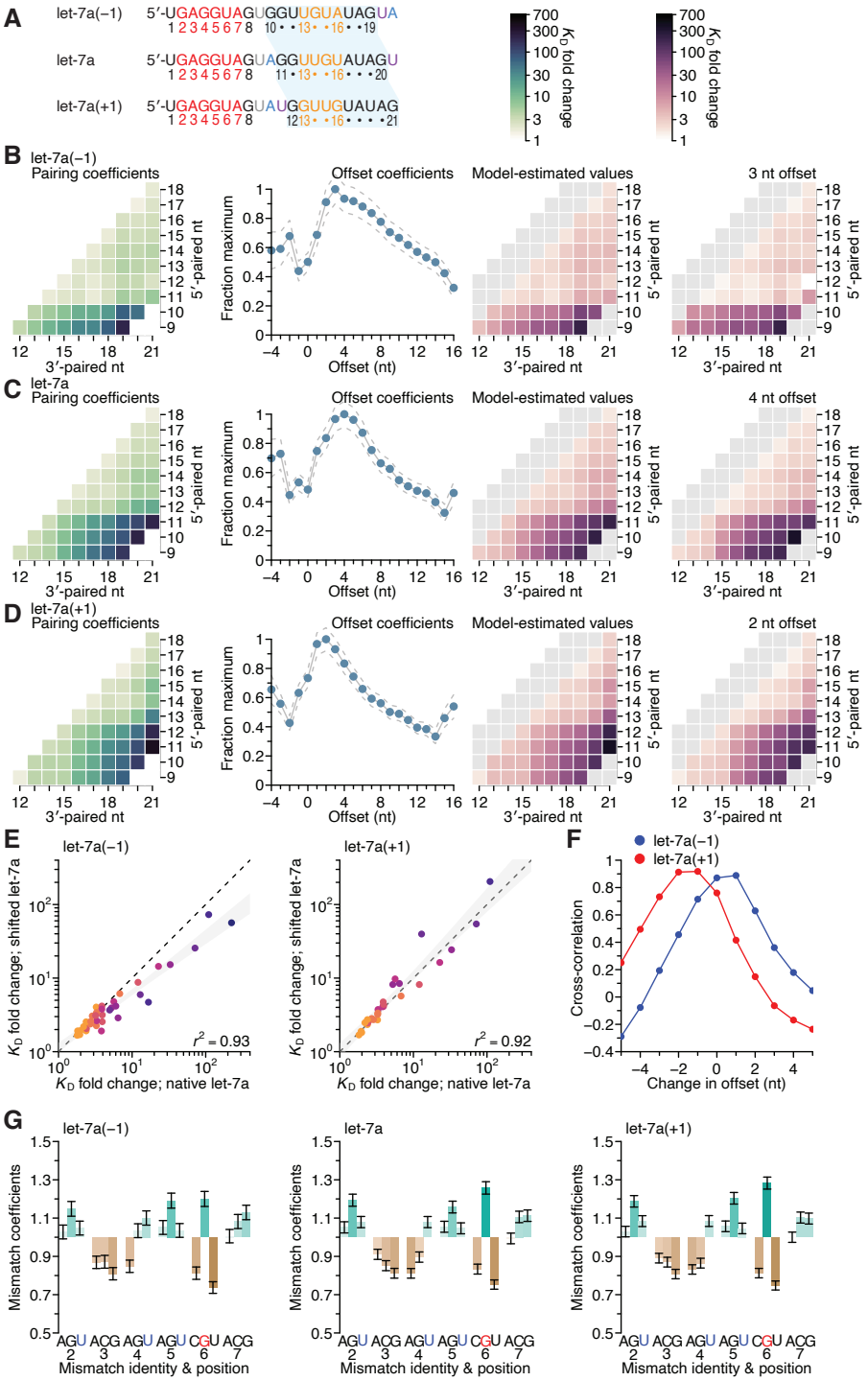


Figure 6. Sequence preferences for 3′ sites are maintained at adjacent positions.
(**A**) Sequences of let-7a(−1), which has a 3′ region permuted one nucleotide toward the 5′ end, native let-7a, and let-7a(+1), which has a 3′ region permuted one nucleotide toward the 3′ end. The 3′ sequence shared between all three miRNAs is shaded in blue, and the A and U nucleotides that were rearranged to generate the permuted variants are in blue and purple, respectively. (**B**–**D**) Pairing and offset coefficients describing the 3′-compensatory pairing of let-7a(−1) (B), let-7a (C, redrawn from Figure 4A, for comparison), and let-7a(+1) (D). Otherwise, this panel is as in

Figure 4A. (**E**) Comparison of the pairing coefficients determined for either let-7a(−1) (left) or let-7a(+1) (right) with those of let-7a. Points are colored according to pairing length, as in Figure 2A. For each graph, the gray region indicates the 95% confidence interval of the least-squares fit to the data (r^2, coefficient of determination), and the dashed line represents $y = x$. (**F**) Cross-correlations of offset coefficients for either let-7a(−1) (blue) or let-7a(+1) (red) with respect to those of let-7a (B and C, middle-left), plotted as a function of the difference in offset coefficients. (**G**) Effects of seed mismatches on 3′-pairing affinities of let-7a(−1) (left), let-7a (middle, redrawn from Figure 4F, for comparison), and let-7a(+1) (right). Otherwise, this panel is as in Figure 4F.

The offset preference of let-7a(+1) shifted between −1 and −2 nt with respect to that of let-7a (Figure 6F), supporting the idea that fewer nucleotides were actually required to bridge the seed and 3′ pairing when the 3′ pairing started at position 12 rather than position 11. The shifted offset preference of let-7a(−1) was between 0 and +1 nt, indicating that a length of 6 nt was nearly equally preferred when the G was at position 10 or 11. Considered in the context of the reduced efficacy of 3′ sites for let-7a(−1), this might indicate that additional bridging nucleotides of the target RNA cannot make up for the reduced benefit of starting pairing at position 10. Finally, the seed-mismatch preferences of both let-7a derivatives were nearly identical to those of native let-7a [Figures 6G; $r^2 = 0.91$ and 0.99 for let-7a(−1) and let-7a(+1), respectively]. Considered together, these results provided further evidence of the independent effects of the seed and 3′ region on 3′ pairing, with the behavior of the 3′ region depending on both sequence and position, with sequence preferences transferable to nearby positions, especially if compensating changes optimize the length of the target segment bridging the seed and 3′ site.

Effects of mismatches within 3′ sites are consistent across miRNAs but explained poorly by the nearest-neighbor model

Having systematically analyzed the contributions of seed-mismatch identity and of the length, position, and offset of perfect 3′ pairing, we next sought to measure the effects of any

imperfections—i.e., mismatches, wobbles, or bulged nucleotides—within this 3′ pairing. Accordingly, we measured the affinities of variants of each site considered thus far, looking at each possible variant that had one of the eight possible imperfections at one position within the site. These eight imperfections considered at each position of interest included three possible mismatched nucleotides (including G:U wobbles), four possible single-nucleotide bulges (occurring opposite the linkage of two miRNA positions and assigned to the more 3′ miRNA position), and one single-nucleotide deletion (i.e., a bulged nucleotide in the miRNA). Consideration of these variants together with the original sites with perfect contiguous pairing resulted in the measurement of K_D values for 38,108, 44,190, and 52,166 sites for let-7a, miR-1, and miR-155, respectively. Incorporating an imperfection invariably reduced affinity of the 3′ site, which indicated that there were no positions at which the altered helical geometry of a mismatch could compensate for its lack of Watson–Crick pairing. Inspection of the effect of each imperfection at each position of the top site of each length revealed that neither bulges nor deletions were characteristically worse for 3′ pairing than were mismatches, and that bulges were not on average worse than deletions (Figures 7A–7C and S14A–S14C). When comparing effects of internal mismatches to those of mismatches occurring at the end of the pairing, no striking differences were observed. Nonetheless, effects at some internal positions were more striking than others, with larger effects observed for mismatches at nucleotides 11 or 12 of let-7a (Figure 7A), at nucleotide 14 of miR-1 (Figure 7B), and between nucleotides 14 and 22 of miR-155 (Figure 7C), which concurred with the importance of extending pairing to G_{11}, G_{12}, and $G_{19}G_{20}G_{21}G_{22}$ of the respective miRNAs.

To investigate mismatch tolerance across the range of miRNA 3′-end positions, we calculated the geometric mean of the K_D fold change for a mismatch at each position for all three miRNAs, averaging both over the three mismatches at each position and over each of the 10-bp

sites that contained the position (Figure 7D). As expected, reduced binding affinity tracked with the importance of the positions for 3′ pairing, with greatest effects observed at G_{11} and G_{12} of let-7a (Figure 7D, left-hand bars at each position), the G_{12}–G_{15} of miR-1 (Figure 7D, middle bars), and G_{13} and G_{15}–G_{21} of miR-155 (Figure 7D, right-hand bars). The greater importance of pairing to G_{13} compared to pairing to C_{12} of miR-155 further supported the idea that pairing to G had a greater impact than pairing to C in the miRNA 3′ region. Nonetheless, extending the analyses of mismatches, wobbles, and bulges to the random-sequence RBNS datasets previously acquired for six miRNAs (Figure S15) indicated that disrupting pairing to either C_{13} or C_{15} of the $C_{13}G_{14}C_{15}$ trinucleotide of lsy-6 almost entirely abolished pairing. Thus, in some contexts, pairing to a miRNA C nucleotide can be as important as pairing to a miRNA G nucleotide, and C nucleotides as well as G nucleotides can help define the positions of most consequential pairing. More generally, these results showed that the effect of a mismatch to a particular nucleotide was informed primarily by the overall importance of that miRNA nucleotide (i.e., its nucleotide identity and position within the miRNA 3′ end) for pairing, rather than whether the target nucleotide fell within the middle or terminus of the 3′ site.

To examine a potential benefit of bulges near the 5′ and 3′ ends of 3′ sites, we considered all possible 10-nt sites for all three miRNAs with programmed libraries, and calculated the fold difference in relative K_D observed when comparing a site with a terminal mismatch to that of the site with a corresponding terminal bulged nucleotide (i.e., the site variant in which the target nucleotide following the mismatch can pair to the mismatched miRNA nucleotide). For each miRNA, a small but significant benefit to terminal bulges was observed (Figure 7E, $p = 2.4 \times 10^{-5}$, 1.4×10^{-6}, and 4.5×10^{-4} for let-7a, miR-1, and miR-155, respectively; one-tailed Wilcoxon signed rank test). Thus, an isolated complementary target nucleotide separated from a longer contiguous stretch of pairing can contribute modestly to site affinity.

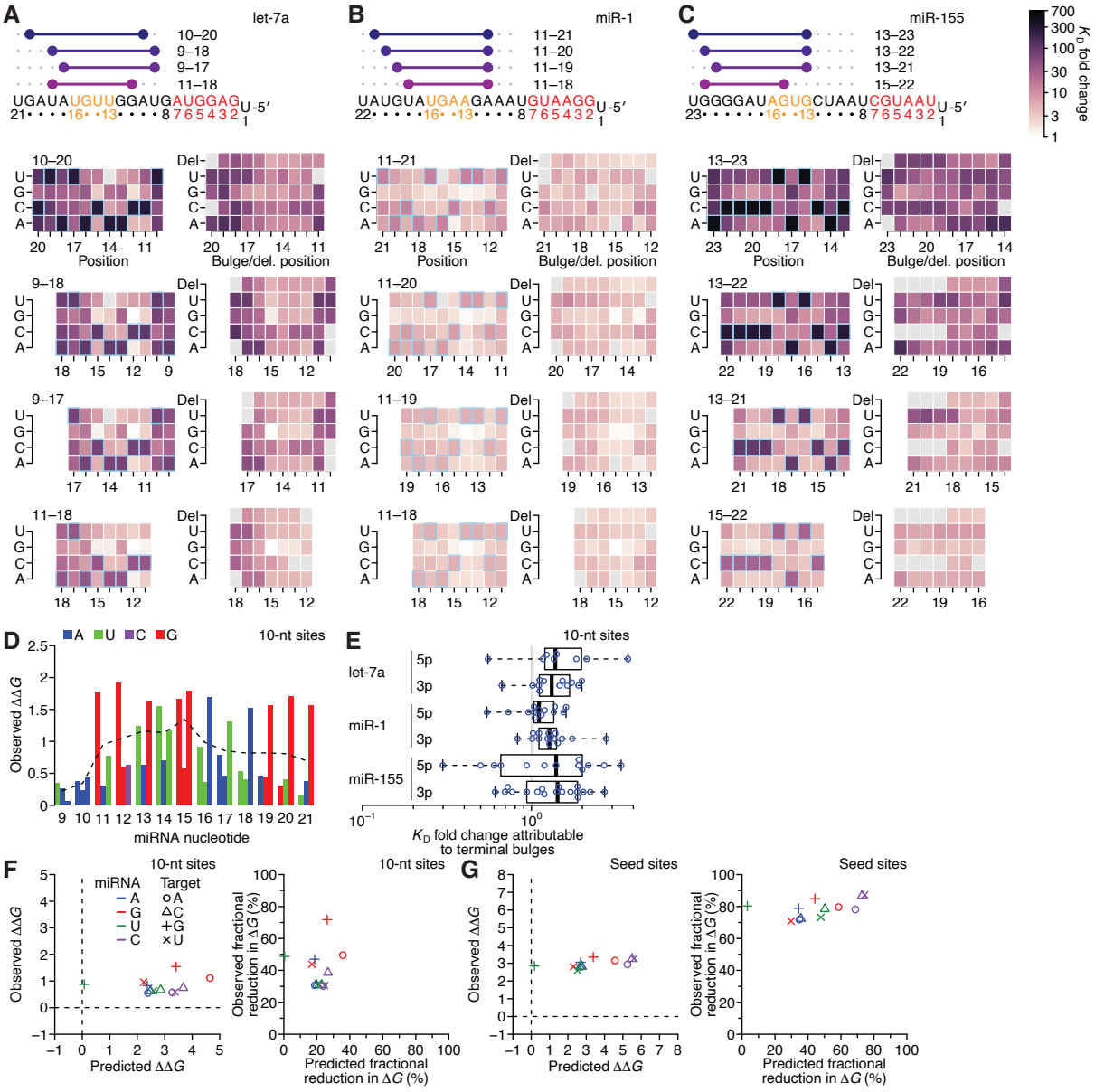


Figure 7. The impact of mismatched, bulged, and deleted target nucleotides on 3′-compensatory pairing.

(**A**) The effect of mismatched, bulged, and deleted target nucleotides on 3′-compensatory pairing to let-7a. At the top is a schematic depicting the highest-affinity 3′-pairing ranges of lengths 8–11 nt, redrawn from Figure 2B. Below, at the left are heat maps corresponding to each of the pairing ranges shown above, indicating the affinities with each of the four possible nucleotides at each position along the site. Cells corresponding to the Waston–Crick match are outlined in blue. Cells for affinities of mismatches that could not be calculated due to sequence similarity to another site type are in gray (e.g., the mismatched U across from position 14, which was indistinguishable from a 6mer-m8 seed site) are in gray. To the right are heat maps that correspond to the same pairing ranges but indicate the effects of an added bulged or a deleted (del.) 3′-target nucleotide. A bulged nucleotide at position *n* corresponded to an extra target nucleotide inserted between the nucleotides pairing to miRNA positions *n* – 1 and *n*. (**B**) The

effects of mismatched, bulged, and deleted target nucleotides for miR-1. Otherwise, this panel is as in A. (**C**) The effects of mismatched, bulged, and deleted target nucleotides for miR-155. Otherwise, this panel is as in A. (**D**) Profiles of 3′-pairing mismatch tolerances. Each bar represents the $\Delta\Delta G$ value when averaging over the three possible mismatches at that position. At each position, the results for let-7a, miR-1, and miR-155, respectively, are plotted as a triplet. Each of the mismatch ΔΔG values was itself an average of the values observed in the context of each 10-nt 3′ site that included the position. The dashed line indicates the average over all three miRNAs, and the color indicates whether the miRNA nucleotide was an A (blue), U (green), C (purple), or G (red). (**E**) The tolerance of bulged nucleotides near the ends of 3′ sites. Plotted are ratios of K_D fold-changes comparing a site that has a bulged nucleotide between the penultimate and terminal base pairs with a site that does not have the terminal base pair (in which case, the bulged nucleotide in the former pairing architecture becomes a terminal mismatch). The box plots indicate the minimum, lower quartile, median, upper quartile, and maximum values. The vertical gray line indicates a K_D fold-change ratio of 1.0. (**F**) Comparison of the measured mismatch $\Delta\Delta G$ values in 3′ sites with values predicted by nearest-neighbor rules. Left, comparison of the average measured ΔΔG value with the average predicted value for each of the 12 possible miRNA–target mismatch combinations. Right, comparison of measured and predicted average fractional reduction in ΔG attributed to each mismatch. The fractional reduction was given by $(\Delta G_{WC} - \Delta G_{mm})/\Delta G_{WC}$, where ΔG_{WC} corresponds to the ΔG of the site with full Watson–Crick pairing, and ΔG_{mm} corresponds to the ΔG of a site containing the mismatch. These average values were calculated using K_D fold-change values determined for 10-nt sites, first averaging results for same position over all 10-nt sites that included the position, then averaging results for that mismatch across all positions of the miRNA that had that mismatch, and then averaging the results across all three miRNAs. Colors and symbols indicate miRNA and target nucleotide identities, respectively (key). (**G**) Comparison of the measured seed-mismatch ΔΔG values with values predicted by nearest-neighbor rules. For each mismatch type, both the measured and predicted $\Delta\Delta G$ values were the average over all occurrences within positions 2–7 for let-7a, miR-1, miR-155, miR-124, lsy-6, and miR-7, using K_D fold changes from analyses of random-sequence AGO-RBNS results. Otherwise, this panel is as in (F).

Next, we calculated the $\Delta\Delta G$ of each mismatch in the context of all 10-nt 3′ sites of the three miRNAs. We first averaged these values over all the contiguous sites, and then over all positions with the same miRNA nucleotide, and then over the three miRNAs, resulting in one global average $\Delta\Delta G$ value for each of the 12 possible miRNA–target mismatch possibilities. Comparison of these values and those predicted using the nearest-neighbor parameters revealed that the effects of the mismatches were typically much lower than expected for free RNA in solution, with no strong relationship between the observed and predicted $\Delta\Delta G$ values (Figure 7F, left; $r^2 = 0.02$). The outlier in this analysis was the miRNA–target U:G wobble, which was as

disruptive as the typical mismatch but predicted to be much less so (Figure 7F, left, green +). Next, to account directly for the reduced binding energy of the fully complementary sites in comparison to their predicted Δ*G* values, we compared the average observed and predicted fractional reduction in Δ*G* of each site caused by each of the twelve mismatch values (Figure 7F, right). For eight of 12 mismatches, the fractional reduction in Δ*G* was within 10% of its prediction, but the miRNA–target A:G, G:G, G:U, and U:G mismatches respectively caused 31%, 42%, 21%, and 48% more reduction in binding energy than predicted. These results indicated that the nearest-neighbor parameters were not suited for predicting the contribution of miRNA 3′ pairing in three respects: 1) the overall contribution to binding energy was far less than that predicted, 2) mismatched target G nucleotides were relatively more deleterious than predicted, and 3) wobble pairing was relatively less favorable than predicted. Indeed, the U:G possibility, which both contained a target G nucleotide and was a wobble, was the mismatch with the greatest deviation from expectation.

For comparison, we repeated these analyses for mismatches within pairing to the miRNA seed (i.e., miRNA positions 2–7), calculating the average ΔΔ*G* and the fractional reduction in Δ*G* for each type of mismatch within pairing to each of the six miRNAs for which there was random-sequence RBNS data (McGeary et al., 2019) (Figure 7G). These analyses indicated that the effects of mismatches within seed pairing also did not agree with predicted pairing energetics, albeit differently than the effects of mismatches within the 3′ pairing. First, a mismatch within the seed pairing had a much larger influence on ΔΔ*G* than did a mismatch within the 3′ pairing. Moreover, the reductions in binding affinities for mismatches within the seed pairing were even more regular than those for mismatches within the 3′ pairing, with a ~3 kcal/mol detriment for each of the 12 mismatch/wobble possibilities (Figure 7G, left). The fractional reduction in Δ*G* had a similarly large and uniform effect size, with no subset of the

mismatch possibilities showing a relationship with that predicted (Figure 7G, right). Thus, the binding preferences at both the seed and 3′ regions of the miRNA were not well characterized by nearest-neighbor rules, although the nature of the deviations differed in these two regions.

Discussion

An Argonaute-loaded miRNA can be divided into three regions: the seed region (nucleotides 2–8), the central region (nucleotides 9–10 or 9–11), and the 3′ region (Figure 1) (Bartel, 2018). Because the most effective 3′ pairing is reported to center on nucleotides 13–16 (Grimson et al., 2007), some subdivide the 3′ region into the 3′-supplementary region (nucleotides 13–16), and the tail (nucleotides 17 to the terminus), while expanding the central region to include nucleotide 12 (Salomon et al., 2015; Sheu-Gruttadauria et al., 2019a; Wee et al., 2012). The structure of AGO2–miR-122 bound to a 3′-supplementary site, which shows that miRNA nucleotides 9–11 are not available for pairing due to both helical distortion and inaccessibility caused by residues of the PIWI and L2 loop, seems to support the notion of a 3′-supplementary region at nucleotides 13–16 (Sheu-Gruttadauria et al., 2019a). However, greater affinities are observed with more extended 3′ pairing (Becker et al., 2019; Sheu-Gruttadauria et al., 2019b), and we found that 3′-site affinities nearly always increased as the potential for pairing expanded to include most of the 3′ region—and in the positive-offset binding mode, some of the central region. Thus, productive 3′ pairing can encompass the entire miRNA 3′ region and should not be thought of as limited to a short 3′-supplementary region. Indeed, the study reporting that pairing to nucleotides 13–16 is most effective for supplementing seed pairing uses a model for predicting the efficacy of 3′ pairing that rewards extension of that pairing into the remainder of the 3′ region (Grimson et al., 2007).

Also problematic for the notion of a short 3′-supplementary region common to all miRNAs was our observation that the positions most important for 3′ pairing differed between different miRNAs. For example, at their optimal offsets, both let-7a and miR-124 preferred pairing to nucleotides 11–14 over pairing to nucleotides 13–16 (Figures 2B, 4A, S8A, and S8D), and the synthetic let-7a(−1) preferred pairing to nucleotides 10–13 over pairing to nucleotides 13–16 (Figure 6B). Moreover, although miR-155 preferred pairing to nucleotides 13–16 over other 4-nt possibilities, when examining 7-nt 3′ sites, it preferred pairing to nucleotides 15–21 over sites that included pairing to nucleotides 13–16 (Figures 3D and 4B). These observations showing that the preferred positions of 3′ pairing can vary so widely between miRNAs, to include virtually any nucleotide downstream of the seed, argued strongly against assigning the same short 3′-supplementary region to all miRNAs.

Although our results showed that preferred pairing often did not correspond precisely to positions 13–16, preferred pairing did always at least partially overlap this segment. Moreover, as pairing lengths increased from 4 to 6 bp, overlap between preferred pairing and this segment increased, such that the preferred 6-nt sites for let-7a, miR-1, miR-155, miR-124, miR-7 and lsy-6 each included pairing to miRNA nucleotides 13–16. The only exception we observed was the preferred 6-nt site for synthetic let-7a(−1), which paired to nucleotides 10–15. Thus, our results explain why an overall preference for pairing to nucleotides 13–16 was detected in meta-analyses of both functional data for 11 miRNAs as well as evolutionary conservation of sites for 73 miRNA families (Grimson et al., 2007). Our key added insight is that sequence identity in the 3′ region—particularly the placement of stretches of G residues—imparts additional preferences that supplement the positional preferences to specify different optimal regions of 3′ pairing for different miRNAs.

Another key insight is evidence of two distinct 3′-binding modes, observed as different offset preferences of let-7a, miR-124, lsy-6, and miR-7 with and without pairing to nucleotide 11 (Figures 2B, 2C, and S9). In one binding mode, an offset of 0 nt is optimal for 3′ pairing starting at position 12, whereas in the other binding mode, additional nucleotides are required to bridge pairing to positions 10 or 11, resulting in optimal offsets that exceed 0 nt. In a crystal structure of AGO2–miR-122 bound to a 3′-supplementary target that pairs to nucleotides 13–16 with an offset of 0 nt, nucleotide 12 is the first nucleotide available for pairing, whereas pairing to nucleotide 11 is occluded by the central gate (Sheu-Gruttadauria et al., 2019a). We suggest that this structure reflects the conformation of the zero-offset binding mode, as it provides a physical model for why extension of potential pairing from nucleotide 12 to 11 results in almost no increased binding affinity for sites with an offset of 0 nt (Figures S9B, S9D, S9F, and S9G). However, another structure will be required to visualize the positive-offset binding mode that enables optimal pairing to let-7a and miR-124, as well as strong pairing to lsy-6 and miR-7. Genetically identified sites inferred to be utilizing this second binding mode include the two let-7a sites within the 3′ UTR of *C. elegans lin-41*, which both include pairing to nucleotide 11 and an offset of +1 nt, as well as the first lsy-6 site within the 3′ UTR of *C. elegans cog-1*, which includes pairing to nucleotide 11 and an offset of +2 nt. The discovery of these two binding modes required knowledge of the interplay between preferred pairing position and preferred pairing offset, which underscored the utility of obtaining affinity measurements for a large diversity of 3′ sites.

The length of a miRNA can modulate its 3′-pairing affinity, in that a 23-nt derivative of miR-122 has a 3-fold longer dwell time than its 22-nt counterpart (Sheu-Gruttadauria et al., 2019a). Of the miRNAs that we examined, miR-155 and miR-7 were each 23 nt in length, whereas the others were shorter. These two miRNAs had the strongest and the weakest 3′

pairing, respectively. The weak 3′ pairing of miR-7 indicated that although increased miRNA length can sometimes improve 3′ binding affinity, it cannot substitute for other features required for high affinity to the miRNA 3′ region.

Early attempts to either explain targeting efficacy or predict target sites used scores incorporating, among other things, the predicted binding energy between the miRNAs and their proposed targets (Doench and Sharp, 2004; Enright et al., 2003; Krek et al., 2005; Lewis et al., 2003; Rajewsky and Socci, 2004). That these metrics were less useful in identifying consequential 3′ pairing than simpler rubrics scoring only the length and position of complementarity (Grimson et al., 2007) suggests that the parameters derived from interactions of purified RNAs in solution are not directly relevant to miRNAs associated with AGO. The breadth of our affinity measurements provided the ability to assess why such parameters are not as useful. Although high correspondence was observed between the predicted ΔG and measured 3′ pairing affinities (Figure S7A), for miR-1 this relationship nearly disappeared when normalizing for pairing length (Figure 4E). For let-7a and miR-155 a relationship was retained after normalizing for length, but four factors limit the utility of using this relationship for ranking target predictions. The first is the strong effect of position, with pairing to the seed much more consequential than pairing to the 3′ region, and pairing at some positions in the 3′ region more consequential than pairing to others, and much more consequential than pairing to positions 1, 9, and often, 10. The second is the effect of primary sequence, as illustrated by the outsized benefit pairing to the G_{11}, G_{12}, and G_{20} nucleotides of let-7a, miR-1, and miR-155, respectively (Figures S7B and S7C). The third is the poor relationship between the predicted and measured effects of some internal mismatches and wobbles (Figure 7F), and the fourth is a lack of a consistent relationship between predicted ΔG and measured binding affinities between miRNAs (Figure S7A, comparing the slope for miR-1 with that of either let-7a or miR-155).

Comparison of the 3′ regions of the four miRNAs that were more effective at 3′ pairing with those of the two that were not suggested a feature that might have conferred higher 3′-pairing affinity: the presence of two or more adjacent G nucleotides (e.g., the $G_{11}G_{12}$ of both let-7a and miR-124, and the $G_{19}G_{20}G_{21}G_{22}$ of miR-155). Although lsy-6 did not have an oligo(G) stretch, it did have a well-positioned $C_{13}G_{14}C_{15}$ trinucleotide, which together with G_{11} was critical for pairing affinity. When considering all four miRNAs together, as well the lack of any GG, CG, or GC dinucleotides within the 3′ regions of miR-1 or miR-7, we suggest that miRNAs with GG, CG, or GC dinucleotides within positions 13–16 are the ones most likely to participate in productive 3′ pairing, and that pairing that extends to an oligo(G) sequence outside of positions 13–16 will preferentially enhance affinity.

The importance of pairing to miRNA G nucleotides, not C nucleotides (other than the $C_{13}G_{14}C_{15}$ of lsy-6), suggested that a miRNA–target G:C base pair is read out differently than a C:G base pair. Perhaps G nucleotides participate in base-stacking interactions that position or pre-organize the guide strand to favor nucleation of 3′ pairing. Alternatively, the explanation might involve target-site accessibility. Pairing to a C in the miRNA 3′ region would require a G in the vicinity of the seed match, which compared to a C would cause poorer target-site accessibility (McGeary et al., 2019), thereby reducing the net contribution to binding.

Our results also revealed a functional difference between 3′-supplementary and 3′-compensatory pairing. The affinity of a 3′ site was relatively constant when it supplemented different sites that had seed matches (Figures 4H and S13), whereas it varied in the context of different 3′-compensatory sites that had different seed mismatches (Figures 4F and S10–S12). The effects of seed mismatches were miRNA-specific and unrelated to their binding affinities (Figure 4G). Additionally, our experiments using chimeric miRNAs demonstrated the

separability of the mismatch effects from the length, position, offset, and nucleotide-identity preferences of the 3′ region (Figure 5).

Pairing to the miRNA 3′ region not only increases site affinity and target repression, but it can also influence the stability of the miRNA itself, in a process called target-directed miRNA degradation (TDMD) (Ameres et al., 2010; Bitetti et al., 2018; Cazalla et al., 2010; Kleaveland et al., 2018; Mata et al., 2015). The handful of target sites known to trigger TDMD have diverse 3′-pairing architectures. For example, degradation of miR-7 triggered by the cellular Cyrano transcript occurs through a canonical 8mer site supplemented with a 3′ site with 14 contiguous pairs to the 3′ end of the miRNA (Kleaveland et al., 2018), whereas degradation of miR-27a triggered by the m169 RNA from murine cytomegalovirus occurs through a canonical 8mer site supplemented with a 3′ site with only six contiguous pairs to the 3′ end of the miRNA. Our finding that that miR-7 has the weakest 3′ pairing among the six miRNAs we studied provides a potential explanation as to why its TDMD trigger Cyrano has such a long 3′ site.

The crystal structures of several known TDMD substrates bound to their corresponding TDMD-inducing target sites reveal a distinct conformation for these AGO–miRNA–target RNA ternary complexes in comparison to ternary complexes that have supplementary pairing involving only nucleotides 13–16 (Sheu-Gruttadauria et al., 2019a, 2019b). During TDMD, this distinct conformation is thought to be recognized by the ZSWIM8 E3 ubiquitin ligase, causing AGO proteolysis through the ubiquitin–proteasome system, which exposes the miRNA to degradation by cellular nucleases (Han et al., 2020; Shi et al., 2020). Our discovery of the two 3′ binding modes raises the question of whether one of them might be more compatible with TDMD, perhaps due to a preference of the ZSWIM8 E3 ligase. Although the TDMD ternary complexes of the published structures all have 3′ pairing beginning at nucleotide 12 or later and offsets of 0 or −1 nt (Sheu-Gruttadauria et al., 2019b) and thereby represent the zero-offset

binding mode, the 3′ pairing between miR-7 and Cyrano begins at G_{11} and has a +2-nt offset, which represents the positive-offset binding mode. Thus, the two 3′ binding modes both appear to enable the miRNA 3′ region to participate in either of its two critical gene-regulatory processes—TDMD and miRNA-mediated repression.

Materials and methods

Human Embryonic Kidney 293T (HEK293T) Cells

HEK293T cells were cultured in DMEM (VWR) with 10% fetal bovine serum (Clonetech) at 37°C with 5% CO_2, and split every third day at ~90% confluency.

Purification of AGO2–miRNA complexes

AGO2–miRNA complexes were generated and purified as described previously (McGeary et al., 2019).

Preparation of programmed RNA libraries

For each of let-7a, miR-1 and miR-155, programmed libraries were constructed by performing in vitro transcription with multiple chemically synthesized DNA libraries, which were then mixed after gel purification. Each library contained 25 nucleotides of entirely randomized sequence, followed by an 8-nt programmed site, followed by either 5 nucleotides of random sequence, in the case of the let-7a and miR-1 programmed libraries, or 4 nucleotides, in the case of the miR-155 programmed libraries. When mixing the programmed library for every experiment other than that with native miR-155, the final programmed library was made by mixing six different libraries, where each of the six libraries contained an 8mer at the programmed site containing a mismatch at one of the six seed positions. In the case of native miR-155, the programmed library

was assembled by mixing the six possible 8mer-mismatch libraries as well as a 6mer-containing library, in which each molecule had all either a C, G, or U at positions 1 and of the programmed site.

Each in dividual library was commercially synthesized (IDT), transcribed, and purified as described previously (McGeary et al., 2019), and then mixed according to the specifications above. The final percentages of the 18 mismatch libraries would be expected to be ~5.6%. The fraction of reads associated with each of the 18 mismatch sites, as measured by sequencing of the input library during each experiment, was 3.4–8.0% for let-7a rep. 1, 2.9–8.7% for let-7a rep. 2, 3.3–7.8% for miR-1, 2.2–6.3% for miR-155, 3.4–8.0% for let-7a(+1) and let-7a(−1), 3.3–7.6% for let-7a–miR-155, and 2.6–6.1% for miR-155–let-7a.

AGO-RBNS

AGO-RBNS was performed as described previously (McGeary et al., 2019).

Analysis of *k*-mer enrichments

Positional enrichments of all 8-nt *k*-mers were calculated by comparison of the sequenced binding sample containing 840 pM AGO2–let-7a complex and 100 nM let-7a–specific programmed library to that of the directly sequenced input library. For each of the two libraries, reads that contained one of the 18 possible 8mer mismatch sites in the correct position (such that the CUACCUCA 8mer-consensus sequence spans positions 26–33 of the read), but did not contain a canonical 8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-A1, or 6mer-m8 site, were used to enumerate all possible 8-nt *k*-mers at each position within the library. Both count tables were normalized such that they summed to 1, and the normalized count table corresponding to the bound sample was divided by that of the input library to arrive at the enrichment of each *k*-mer at

each position within the library. These *k*-mers were ranked according to the sum of the top five positional enrichments of each, considering positions 9–26 of the library, with positions 9 and 26 referring to the 3′-most (i.e., abutting the programmed site), and 5′-most (i.e., abutting the 5′ constant region) positions, respectively.

Read assignment of miRNA sites with contiguous 3′-pairing for the programmed-libraries experiments

When counting seed sites and fully complementary 3′ sites within the programmed libraries, individual reads were first queried for whether they did or did not include a canonical or 8mer-mismatch site at the programmed region (i.e., at positions 26–33). Reads containing a canonical site despite their not having been included within the programmed-library design (e.g., an 8mer or 7mer-m8 site) were still counted, but their measured relative K_D values would are not considered in this study, owing to the ambiguity of whether the error took place during chemical synthesis, *in vitro* RNA transcription, library preparation, or Illumina sequencing.

Those reads containing a seed site at the programmed region were further assigned to one of four categories: 1) reads containing neither a seed site (defined as an 8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-m8, or 6mer-A1 site, or one of the 18 possible canonical 8mer sites with a mismatch within positions 2–7) nor a 3′ site (defined as a site of 4–11 bp of contiguous complementarity to a region of the miRNA spanning position 9 to the 3′-most nucleotide), 2) reads containing at least one seed site but no 3′ sites, 3) reads containing at least one 3′ site but no seed sites, and 4) reads containing at least one seed site and at least one 3′ site. This categorization was chosen in order to assess the contribution of each subsequence of the 3′ end only using reads with little seed-binding capacity other than at the programmed region.

The categorization of each read proceeded through the following steps: 1) any seed sites in addition to that of the programmed site were identified, looking within the entire random-programmed-random region with three nucleotides of constant sequence appended to the 5′ and 3′ ends of the read, 2) the 28-nt segment was queried for the longest contiguous match to the miRNA 3′ end, retaining multiple putative sites in the case ties. Any putative site or sites 4–11 nt in length that were not contained within any of the seed sites within the read (if such sites were present) were counted as 3′ sites, and 3) the read was then assigned one of the four categories described above. In the case of reads with only seed sites or only 3′ sites, the read count was split between each of these sites, recording the type of site, the identity of the programmed site, and the distance between the two. In the case of reads with both seed sites and 3′ sites, the read was split between all seed-and-3′-site pairs, recording the names of the seed, 3′, and programmed site for each.

Analysis of the read data in this way yielded tables of counts associated with categories of reads with 1) only programmed sites, 2) seed-and-programmed site pairs with positional information, 3) 3′-and-programmed site pairs with positional information, 4) seed-and-3′-and-programmed site triples with no positional information, and 5) reads without a correct mismatch site. These count tables were either used directly for relative K_D estimation, or first combined with respect to the identity of their programmed sites prior to relative K_D estimation. In this case, all counts corresponding to reads with identical site and positional information were summed into two categories: those whose programmed site was an 8mer-mismatch site, and those whose programmed site was one of the canonical sites.

Read assignment of miRNA sites with contiguous 3′-pairing for the random-library experiments

When counting seed sites and fully complementary 3′ sites within the random-sequence libraries, the 37-nt random-sequence region of each read was appended with 3 nt of constant sequence at either end, except in the case of miR-1, for which the 5′-most 36 nt of the random-sequence region was appended with 3 nt of only the 5′ constant sequence, due to the sequence bias present at the very 3′ end of these libraries caused by erroneous lack of a TCG sequence in the 3′ constant region required for pairing to the Illumina reverse-primer sequence during bridge-amplification (McGeary et al., 2019). The relevant portion of each read was queried for all seed sites (defined as above) and all 3′-sites between 4–11 nt in length, allowing individual seed sites to overlap, and individual 3′ sites to overlap. If the read contained only seed sites, or only 3′ sites, the read counts were split evenly between each site found within. If a read contained at least one seed site and at least one 3′ site, each 3′ site was checked for any amount of overlap with any seed sites. If the 3′ site overlapped a seed site with the 3′ site being the 5′-most site, the 3′ site was trimmed to not include the region overlapping the seed site. If the trimming the 3′ site did not result in its being <4 nt in length, it was putatively retained. Any 3′ sites that either overlapped any seed sites from the 3′ end, were entirely contained within a seed site, or were entirely contained a seed site, were discarded.

If one or more 3′ site persisted for the read after querying for any seed-site overlap, all the 3′ site of length equal to the longest 3′ site were retained. All possible bipartite sites associated with that read were then enumerated, in which each seed site was considered to form a bipartite site with all 3′ sites that were 5′ of that seed site. The read was then split among any bipartite sites identified. In the event that no bipartite sites were identified, the read was split among all the seed and 3′ sites equally. While this procedure ensures the equal partitioning of read counts in the case of multiple seed and 3′ sequence elements within a given read, in practice only a

small fraction of reads contained multiple seed sites and a 3′ site, or a seed site and multiple 3′ sites.

This yielded tables of counts associated with categories of reads with 1) only seed sites, 2) only 3′ sites 3) single seed-and-3′ bipartite sites with recorded inter-site spacing, and 4) neither a seed nor 3′ site (referred to as "no site"). These count tables were either used directly for relative K_D estimation, or the bipartite sites were combined with respect to the identify of their seed sites prior to relative K_D estimation. In this case, all counts corresponding to reads with the same 3′ site and distance from the miRNA position 8 of the seed site were summed into two categories: those whose seed site was an 8mer mismatch site, and those whose seed site was one of the canonical sites.

Relative K_D assignment

Relative K_D assignment was performed as described previously (McGeary et al., 2019).

Correction of programmed library experiment–derived relative K_D values using data from random-library experiments

Due to the deviation from the expected linear relationship between the relative K_D values calculated for seed sites (defined as the 8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-m8, and 6mer-A1 sites) and 8mer-mismatch sites (Figures S1B–S1D, left) we applied locally estimated scatterplot smoothing (LOESS) to generate an empirical correction to apply to this data, as has been used previously for correction of mRNA abundance in metabolic labeling experiments as a function of uridine content (Schwanhäusser et al., 2011). We note that the relative K_D values of the seed sites for the programmed-library experiments used for this correction were derived from geometric mean of the relative K_D values of each site measured at each position within the library and in the

context of each of the 18 8mer-mismatch sites, while the relative K_D values of the 8mer-mismatch sites were derived from the reads associated their occurrence in the programmed sites in the absence of any seed sites or 3′ sites 4–11 nt in length. The K_D values of these sites for the random-library experiments were derived from counts corresponding to single instances of these sites within the reads.

We corrected the programmed-library relative K_D values by calculating R_i, defined as:

$$R_i = \ln \frac{K_{r,i}}{K_{p,i}} , \tag{3.1}$$

where $K_{r,i}$ and $K_{p,i}$ refer to the relative K_D values derived from the random-library and programmed-library experiments, respectively, for each site i. LOESS was used to fit a nonlinear function describing R_i as a function of $K_{p,i}$:

$$R(K_p) \sim f_{LOESS}(x = \ln K_p) . \tag{3.2}$$

This function was then used to correct each programmed library–derived relative K_D value by multiplying each value by the output of the function with itself as input:

$$K'_{p,i} \equiv K_{p,i} \times e^{f_{LOESS}(x = \ln K_{p,i})} . \tag{3.3}$$

These transformed $K'_{p,i}$ values were used throughout the study other than in the left-hand panels of Figures S1B–S1D. LOESS was implemented in *R* using the *loess* function as part of the *stats* package, with "span" and "surface" parameters set to 10 and "direct", respectively.

Re-analysis of data from Becker, Ober-Reynolds *et al.* (2019)

The 22,300 K_D values measured for let-7a were analyzed using the table provided as supplementary data (Becker et al., 2019). For each of the target sequence–and–K_D value pair, the target sequence was queried for any of canonical or 8mer-mismatch sites, hierarchically looking

first for the 8mer, any 8mer-mismatch sites, any 7mer-m8 or 7mer-A1 sites, or any 6mer, 6mer-m8, or 6mer-A1 sites. If one or more of these sites were found, the target sequence 5′ of each site was queried for its longest stretch of complementarity to the 3′ end of let-7a (i.e., all nucleotides 3′ of position 9), and if it were between 4–11 nt in length, the seed site, the 3′ site, and intervening nucleotide length would be ascribed to that target sequence. If there were multiple 3′ sites of the same, longest length were present 5′ of the seed site, both 3′ sites were ascribed to the site. Upon using the sequences to define the bipartite site information for each target RNA, only those target RNAs with a single bipartite site, or a single seed site and no 3′ pairing between 4–11 nt in length, were included in the downstream analyses. From these sites, we calculated the K_D fold change for each available 3′ site, offset, and seed site combination (Figure S3) by dividing the geometric mean of the K_D values of target RNAs containing that bipartite site by the geometric mean of that of the target RNAs containing the seed site with no 3′ pairing, except when calculating the K_D fold change values for bipartite sites with the 8mer-xA5 seed site (Figure S3F). Because no target RNAs fit our criteria as containing only the 8mer-xA5 site and no 3′ pairing 4–11 nt in length, we used 10 nM as the reference K_D, which was the lower limit of detection measured, and was the measured K_D for 7 of the 16 8mer-mismatch sites present in the data.

Thermodynamic modeling of miRNA 3′-compensatory pairing binding affinity yielding pairing and offset coefficients

In order to separate the intrinsic pairing preferences of each miRNA 3′ end from the effects of varying the offset of pairing, we fit a thermodynamic model of 3′ binding efficacy to the K_D fold-change values measured when summing the counts from each of the 18 8mer-mismatch sites. The model was constructed to produce a $\log_{10}$-tranformed K_D fold-change values, denoted here

using κ, as a function of the 5′ terminus of the pairing i, the 3′ terminus of pairing j, and the offset between the seed and 3′ pairing k. In order to make no assumptions regarding the thermodynamic nature of 3′-end binding (e.g., that each nucleotide would contribute independently to the binding energy), and as well to make no assumptions about the nature of the offset preferences, the model included two sets of categorical coefficients, one set $\alpha_{i,j}$ describing the 3′ pairing range as a function of the 5′ and 3′ termini of pairing indices i and j, and another set β_k describing the offset preferences as a function of the offset index k:

$$\kappa(i,j,k) = \kappa\left(\alpha_{i,j}, \beta_k\right) . \tag{3.4}$$

Because the nature of the relationship between the pairing range and the offset preferences could not be known *a priori*, we constructed three variants of the model function $\kappa(i,j,k)$:

$$\kappa_a(i,j,k) = \alpha_{i,j} + \beta_k \tag{3.5.1}$$
$$\kappa_m(i,j,k) = \alpha_{i,j}\beta_k \tag{3.5.2}$$
$$\kappa_{mc}(i,j,k) = \alpha_{i,j}\beta_k + \gamma, \tag{3.5.3}$$

where $\kappa_a(i,j,k)$, $\kappa_m(i,j,k)$, and $\kappa_{mc}(i,j,k)$ describe additive, multiplicative, and multiplicative-plus-constant models. We note that an additive-plus-constant variant is trivially equivalent to the additive model, since the constant term can be subsumed by (i.e., added to) either of the $\alpha_{i,j}$ or β_k coefficients.

All the models described by equations (3.5.1)–(3.5.3) were fit to the data by minimizing a cost function giving the summed squared-loss between the measured $\log_{10}$-transformed K_D fold-change values $y_{i,j,k}$ and their corresponding model predictions $\kappa(i,j,k)$ for all pairing-range and offset combinations i, j, and k with 3′-pairing lengths 4–11 nt and offset between −4 and +16 nt:

$$f_{cost,\kappa}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=9}^{n_m-3} \sum_{j=i+3}^{n_m} \sum_{k=-4}^{+16} \left(y_{i,j,k} - \kappa(i,j,k)\right)^2, \tag{3.6}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the vector of all $\alpha_{i,j}$ and all β_k coefficients, respectively, and n_m

represents the length of the miRNA. This cost function was minimized with the *optim* function in R using the L-BFGS-B method, supplying the cost function and its gradient and setting the "maxit" parameter to 1×10^7. When optimizing all three models, all $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ parameters were initialized at 0 and bounded between 0 and 10 during the optimization.

Because the multiplicative model $\kappa_m(i, j, k)$ (r^2 = 0.92, 0.86, and 0.96 for let-7a, miR-1, and miR-155, respectively, Figure S6D) performed significantly better than that of the additive model $\kappa_a(i, j, k)$ (r^2 = 0.81, 0.81, and 0.94), and because the multiplicative-plus-constant model $\kappa_{mc}(i, j, k)$ provided only marginally increased performance (r^2 = 0.93, 0.87, and 0.96) while decreasing model interpretability, as the constant term physically corresponded to a benefit to binding irrespective of the manner of the pairing to the miRNA 3′ end, we selected the multiplicative model. For the purposes of interpretation of the model coefficients, we re-scaled the coefficients as follows:

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} \times \max \boldsymbol{\beta} \tag{3.7.1}$$

$$\boldsymbol{\beta}' = \frac{\boldsymbol{\beta}}{\max \boldsymbol{\beta}}\ , \tag{3.7.2}$$

which, because none of the coefficients were negative, caused each offset coefficient β'_k to be between 0 and 1, thereby corresponding to a different fractional reduction in binding energy for each offset *k*. Each re-scaled pairing range coefficient $\alpha'_{i,j}$ therefore also represented the maximum K_D fold change that could be obtained by contiguous pairing to nucleotides *i* through *j*.

We estimated the model error by calculating the asymptotic covariance matrix $\boldsymbol{V}\langle\boldsymbol{\theta}\rangle$, where $\hat{\boldsymbol{\theta}}$ is the vector of all optimal pairing-range and offset coefficients. This is standardly approximated by

$$\boldsymbol{V}\langle\hat{\boldsymbol{\theta}}\rangle = \frac{f_{cost,\kappa}\left(\hat{\boldsymbol{\theta}}\right)}{n - p}\left(\boldsymbol{A}\langle\hat{\boldsymbol{\theta}}\rangle^T \boldsymbol{A}\langle\hat{\boldsymbol{\theta}}\rangle\right)^{-1}. \tag{3.8}$$

where *n* is the total number of data points, *p* is the total number of model parameters (i.e., the

length of vector $\hat{\boldsymbol{\theta}}$), and $\boldsymbol{A}\langle\hat{\boldsymbol{\theta}}\rangle$ represents the matrix of partial derivatives $\frac{\partial\kappa}{\partial\theta}$ (Alper and Gelb, 1990). From the covariance matrix, the 95% confidence intervals for each model coefficient are given by

$$\hat{\boldsymbol{\theta}} \pm t_{\alpha=0.975,\nu=n-p}\sqrt{\mathrm{diag}\left(\boldsymbol{V}\langle\hat{\boldsymbol{\theta}}\rangle\right)}\,, \tag{3.9}$$

where $t_{\alpha=0.975,\nu=n-p}$ represents the t statistic for 97.5% confidence with $n - p$ degrees of freedom. Because the form of the model described allows the cost function to be minimized with an infinite number of distinct solutions (where, given a particular optimal $\hat{\boldsymbol{\theta}}$ comprised of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, any $\widehat{\boldsymbol{\theta}'}$ comprised of $c\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}/c$ is an equivalent solution), the matrix given by $\left(\boldsymbol{A}\langle\hat{\boldsymbol{\theta}}\rangle^{T}\boldsymbol{A}\langle\hat{\boldsymbol{\theta}}\rangle\right)$ is not linearly independent, and thus cannot be inverted as required in equation (3.8). This issue is circumvented by arbitrarily fixing one parameter in the course of the optimization. We therefore optimized the model 21 times, fixing each β_k coefficient at 1 during the optimization, determining the 95% confidence intervals for the all the other coefficients, and then rescaling all parameters. This led to 21 different estimates of the confidence intervals for each pairing coefficient, and 20 distinct estimates of the confidence intervals for each offset coefficient, which were averaged to produce the error estimates reported throughout the study. We note that because the parameters were re-scaled after both the optimization and confidence interval calculation, the final, re-scaled parameter values obtained were identical in each of the 21 optimization routines.

Nearest neighbor rules–based prediction of 3′-compensatory pairing ΔG

For comparison with each pairing-range coefficient beginning at position *i* and ending at position *j*, the predicted ΔG of duplex formation between sequence of the miRNA beginning at position 9 and the sequence reverse-complementary to miRNA positions *i*–*j*, with no non-complementary

nucleotides appended to either terminus, was calculated via RNAduplex, as part of the ViennaRNA package, through its Python interface (Lorenz et al., 2011).

Thermodynamic modeling of binding affinity of miRNA 3′ end yielding pairing, offset, and mismatch coefficients

We extended the thermodynamic model of 3′-end binding efficacy to include seed-mismatch effects, by using as input data the $\log_{10}(K_D$ fold-change) values measured for each of the 18 8mer-mismatch sites separately. This model took the form of:

$$\kappa_2(i, j, k, l) = \alpha_{i,j}\beta_k\delta_l. \tag{3.10}$$

where $\alpha_{i,j}$ and β_k represented the pairing-range and offset preferences as before, and δ_l represented the additional set of 18 seed-mismatch coefficients. The updated cost function was therefore

$$f_{cost,\kappa_2}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=9}^{n_m-3} \sum_{j=i+3}^{n_m} \sum_{k=-4}^{+16} \sum_{l=1}^{18} \left(y_{i,j,k,l} - \kappa_2(i, j, k, l) \right)^2, \tag{3.11}$$

where $\boldsymbol{\delta}$ represents the vector of all δ_l coefficients. The optimization was performed identically to as before, with the γ parameters initialized at 1, and bounded between 0 and +10 during the optimization. After the optimization, the coefficients were re-scaled as

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} \times \max \boldsymbol{\beta} \times \text{mean } \boldsymbol{\delta} \tag{3.12.1}$$

$$\boldsymbol{\beta}' = \frac{\boldsymbol{\beta}}{\max \boldsymbol{\beta}} \tag{3.12.2}$$

$$\boldsymbol{\delta}' = \frac{\boldsymbol{\delta}}{\text{mean } \boldsymbol{\delta}}, \tag{3.12.3}$$

which preserved the same interpretation of each $\alpha'_{i,j}$ and β'_k coefficient as with the prior model, and further parameterized each δ'_l to represent the multiplicative deviation in binding caused by each seed-mismatch type l, with average of all 18 effects set to that of being multiplied by 1.

The 95% confidence intervals of this model also used equations (3.8) and (3.9). However, because this model was the product of three sets of categorical coefficients, one coefficient from each of two sets was required to be fixed while performing the error determination. We therefore optimized the model 21 × 18 times, fixing one β_k coefficient at 1 and one δ_k at 1 during the optimization, determining the 95% confidence intervals for the all the other coefficients, and then rescaling all parameters. This led to 21 × 17 = 357 different estimates of the confidence intervals for each seed-mismatch coefficient, which were averaged to produce the error estimates reported throughout the study.

Empirical assessment of contribution of seed-type to 3′-compensatory and 3′-supplementary pairing using the random-library AGO-RBNS experiments

When analyzing the effects of seed-type on K_D fold change in the random-library experiments, we first attempted to apply the modeling approach as used when analyzing the data for the programmed-library experiments. However, we found that the low numbers of read counts led to significant spareness with respect to all possible pairing range, offset, and seed-mismatch combinations, such that modeling using these data could not be reliably performed. We therefore analyzed the differences between benefit of 3′-supplementary pairing between each of the six canonical sites (8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-m8, and 6mer-A1), and as well a representative 3′-compensatory site given by summing the read counts of all 18 8mer-mismatch sites.

To compare these sites, we first took, for each miRNA, all 3′ pairing-range possibilities of 4 or 5 nt in length and whose 5′ position of pairing was between nucleotides 9 and 18 of the miRNA, and determined for each the offset with the optimal average $\log_{10}(K_D$ fold change) over

the aforementioned site categories, thereby constructing a 7 (site types) × 20 (pairing range) matrix of $\log_{10}(K_D$ fold change) values. This matrix was then sorted column-wise by the average $\log_{10}(K_D$ fold change) of each pairing range, and then this value was subtracted from each column, such that the values within each column reported on the deviation of each site type from the average, for that pairing-and-offset possibility. These deviations were then averaged for each of the seven site types over the top five (i.e., the top quartile) of pairing-and-offset possibilities, to give the empirical contribution of each site type to 3′ binding affinity, in comparison to that of the average. These values are plotted for all six miRNAs in Figure 4H, and the data tables from which they were calculated are visualized in Figure S13, with the columns used for the final averaging indicated.

Read assignment of miRNA sites with 3′-end mismatched, bulged, and deleted nucleotides for the programmed-library experiments

To calculate the effect of all possible mismatched, bulged, and deleted nucleotide on the binding affinity of a particular fully paired 3′ site measured in the course of the programmed-library experiments, the site counting was repeated for each fully paired 3′ site, enumerating these sites only for that particular site. This was done to reduce the total number of sites being counted and subsequently used to calculate K_D, and as well to reduce the possibility of assignment problems owing to any mismatched, bulged, or deleted-nucleotide 3′ sites (hereafter referred to as "imperfect 3′ sites") from one region of the miRNA 3′ end being identical to that of any other region of the miRNA 3′ end.

The site counting was performed similarly to that of the fully paired 3′ sites, with some differences: those reads containing a seed site at the programmed region were still assigned to

one of four categories, with the definition of a 3′ site expanded to include any fully paired 3′ site of length 4–11 nt in length pairing to the miRNA 3′ end in addition to any imperfect 3′ sites derived from the particular fully paired 3′ site. The categorization of each read proceeded through the following steps: 1) Any seed sites in addition to that of the programmed site were identified, looking within the entire random-programmed-random region with 3 nt of constant sequence appended to the 5′ and 3′ ends of the read. 2) The 28-nt segment comprising 3 nt of the 5′-constant sequence and 25 nt of random sequence was queried for any instances of any imperfect 3′ sites, with any deleted- or mismatched-nucleotide sites that were contained with another mismatched-or bulged-nucleotide site not counted (e.g., the let-7a 11mer-m11–20 with a mismatched U at position 20 is inherently contained within the 11-mer-m11–20 with a bulged U opposite position 20, but only the bulged-nucleotide version of the site would be recorded). Any imperfect 3′ sites were also queried to make sure that the nucleotide on either side of the site was not complementary to the next corresponding position of the miRNA guide. We note that if such an imperfect 3′ were found but failed these criteria, any fully paired 3′ sites were not counted toward that read. 3) The 28-nt segment comprising 3 nt of the 5′-constant sequence and 25 nt of random sequence was queried for the longest fully paired 3′ site, retaining multiple putative 3′ sites if multiple were of the longest length. If there were any putative, fully paired 3′ sites >4 nt in length that were not contained within any of the seed sites within the read (if such sites were present), and if any imperfect 3′ sites had also been identified, the length of the fully paired 3′ site or sites was compared to that of the imperfect 3′ sites, and only the category of 3′ site that was longer was retained. If the contiguous 3′ site was longer than the mismatched-, bulged-, or deleted nucleotide sites, these were no longer considered associated with the read. Lastly, if any of the contiguous 3′ sites were ≥11 nt in length, neither the fully paired nor imperfect 3′ sites

were counted. 4.) The read was then assigned one of the four categories, and the read count split as described when assigning contiguous 3′ sites.

The relative K_D values used in Figure 7 were derived from that when summing the counts for all 18 mismatch sites in the programmed region, with the individual values of all of the imperfect 3′ sites corresponding to a particular fully paired site derived from the geometric mean of the three contiguous offset values at which the K_D fold-change of the fully paired site was the greatest.

Read assignment of miRNA sites with 3′-end mismatched, bulged, and deleted nucleotides for the random-sequence experiments

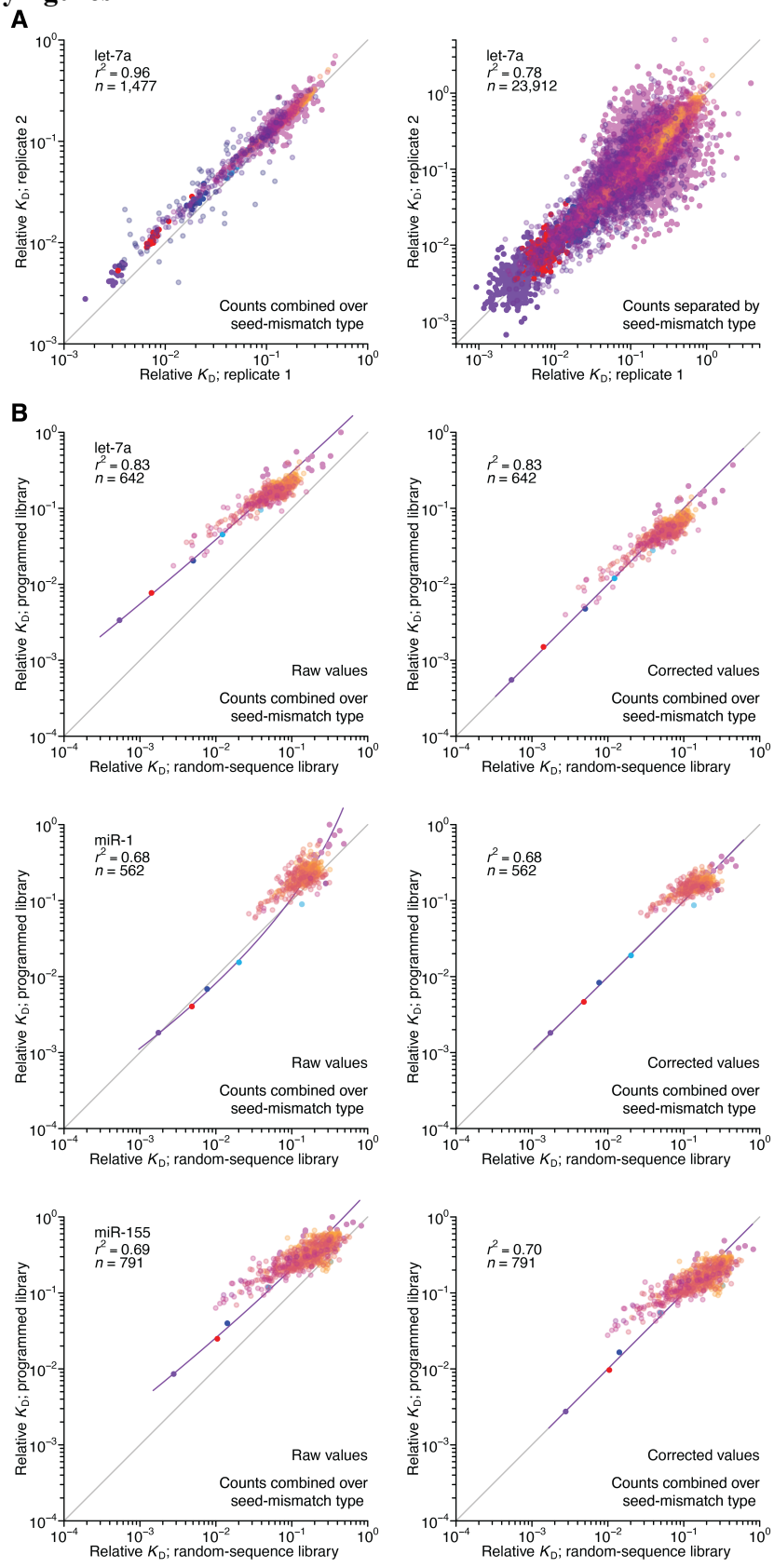
When counting all imperfect 3′ sites within the random-library experiments, we similarly preformed the read counting and relative K_D fitting separately for each fully paired site from which the imperfect 3′ sites were derived. Individual reads were queried for all seed and fully paired 3′ sites as described for the fully paired site counting for the random-library experiments, in addition to being queried for any imperfect 3′ sites derived from a particular fully paired site. If a read contained at least one seed site and at least one fully paired or imperfect 3′ site, each 3′ site was checked for any amount of overlap with any seed sites. If a fully paired 3′ site overlapped a seed site with the 3′ site being the 5′-most site, the 3′ site was trimmed to exclude the region overlapping the seed site, and putatively retained if the site was still ≥4 nt in length. As before, any fully paired 3′ sites that either overlapped any seed sites from the 3′ end, contained a seed site within them, or were entirely contained within a seed site, were discarded. Any imperfect 3′ sites identified were queried to make sure that read positions just outside their

limits were not complementary to the corresponding miRNA positions, and as well that they did not overlap any of the seed sites within the read.

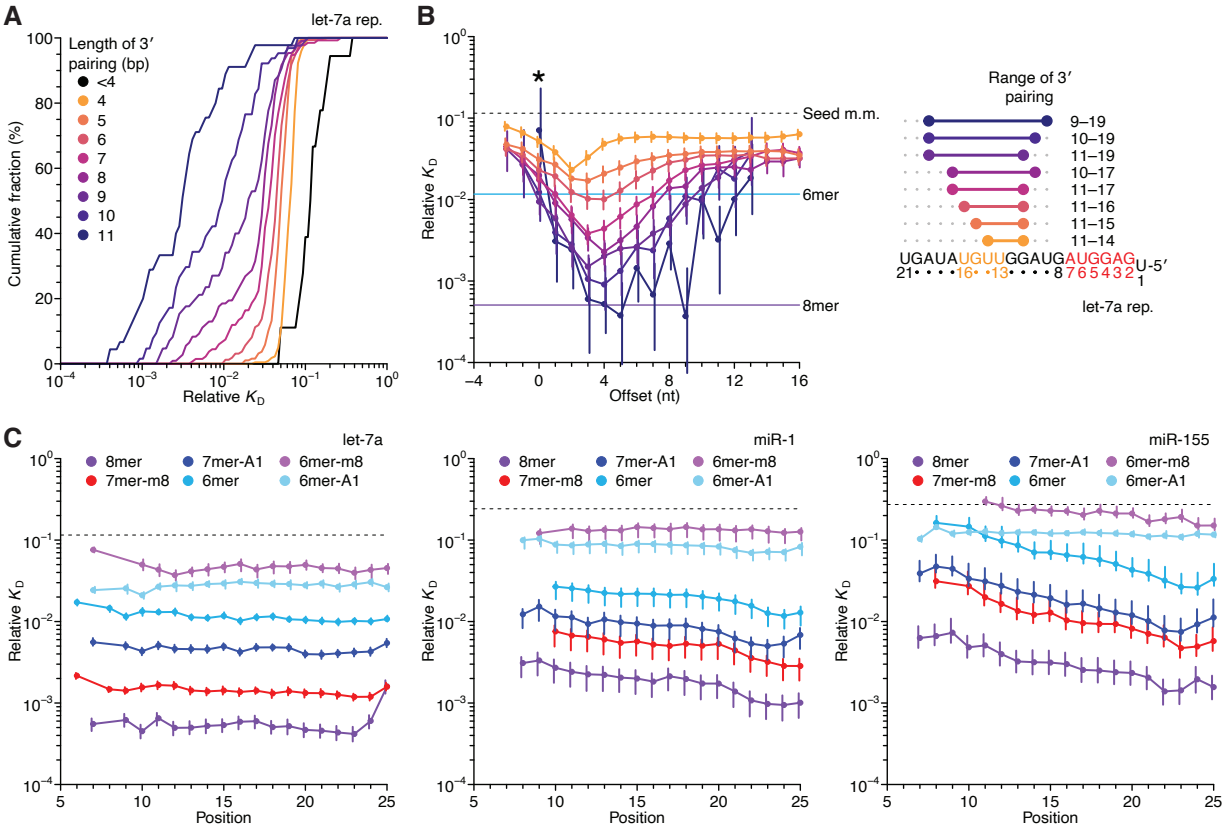
If one or more 3′ site (either fully paired or imperfect) persisted, all of the fully paired 3′ sites of length equal to that of the longest fully paired 3′ site in the read were retained. If at least one fully paired and one imperfect 3′ site persisted, the length of fully paired 3′site or sites was compared to that of the fully paired site from which the imperfect sites were derived, and if it was shorter, the imperfect 3′ site or sites were retained, and the fully paired site or sites discarded. If it were longer, the fully paired 3′ site or sites were retained, and the imperfect site or sites discarded. All possible bipartite sites associated with that read were then enumerated, in which each seed site was considered to form a bipartite site with all 3′ sites that were fully 5′ of that seed site. The read was then split among any bipartite sites identified. In the event that no bipartite sites were identified, the read was split among all the seed and 3′ sites equally. While this procedure ensures the equal partitioning of read counts in the case of multiple seed and 3′ site within a given read, in practice only a small fraction of reads contained multiple seed sites and a 3′ site, or a seed site and multiple 3′ sites.

The relative K_D values shown in Figure S15 were derived from that when summing the counts for all 18 mismatch sites, with the individual values of all of the imperfect sites corresponding to a particular fully paired site derived from the geometric mean of the three contiguous offset values at which the K_D fold-change of the fully paired site was the greatest.

Supplementary figures

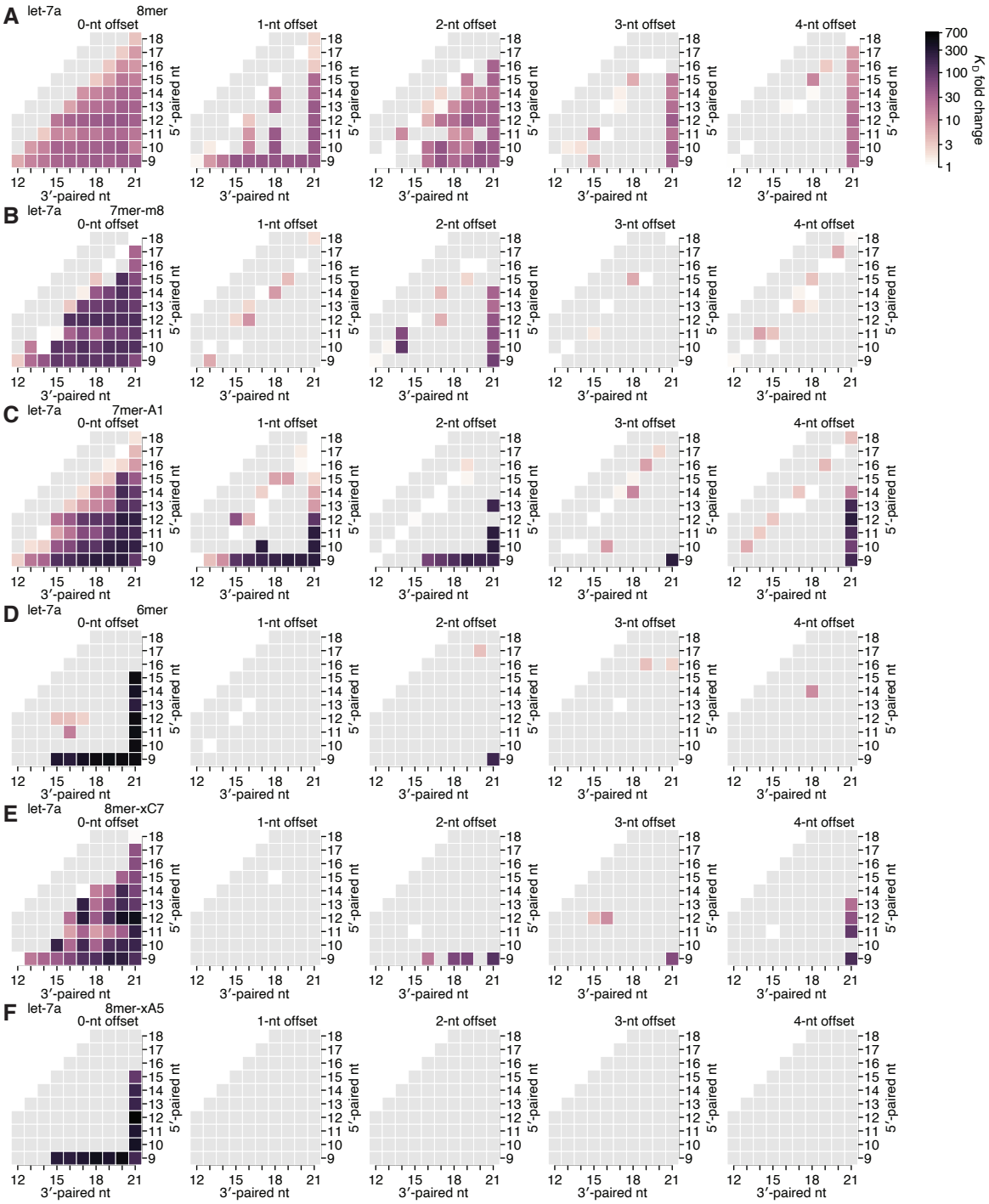


Supplemental Figure 1. Reproducibility of AGO-RBNS with programmed libraries and correspondence with random libraries.
(**A**) Pairwise comparison of replicate relative K_D values measured using the let-7-programmed library and AGO2–let-7a when combining (left) or separating (right) reads based on the identity of the seed-mismatch site at the programmed region of the library. Each of the two replicate experiments was performed with independent preparations of both the library and the purified AGO–miRNA complex. The K_D values correspond to both seed sites and 3′-compensatory sites spanning 4 (orange) to 11 (dark blue) contiguous base pairs in length. The r^2 value reports on the coefficient of determination between the log-transformed relative values of each replicate. (**B**) Pairwise comparison of the relative K_D values measured for let-7a (top row), miR-1 (middle row), and miR-155 (bottom row) in the programmed-library experiments to that of the random-library experiments, prior to (left-hand column) and after (right-hand column) correction of each of the programmed library–derived measurements using the random-library experiments using LOESS. Otherwise, this panel is as in (A).



Supplemental Figure 2. Further analysis of replicate let-7a experiment and positional enrichment of canonical sites.

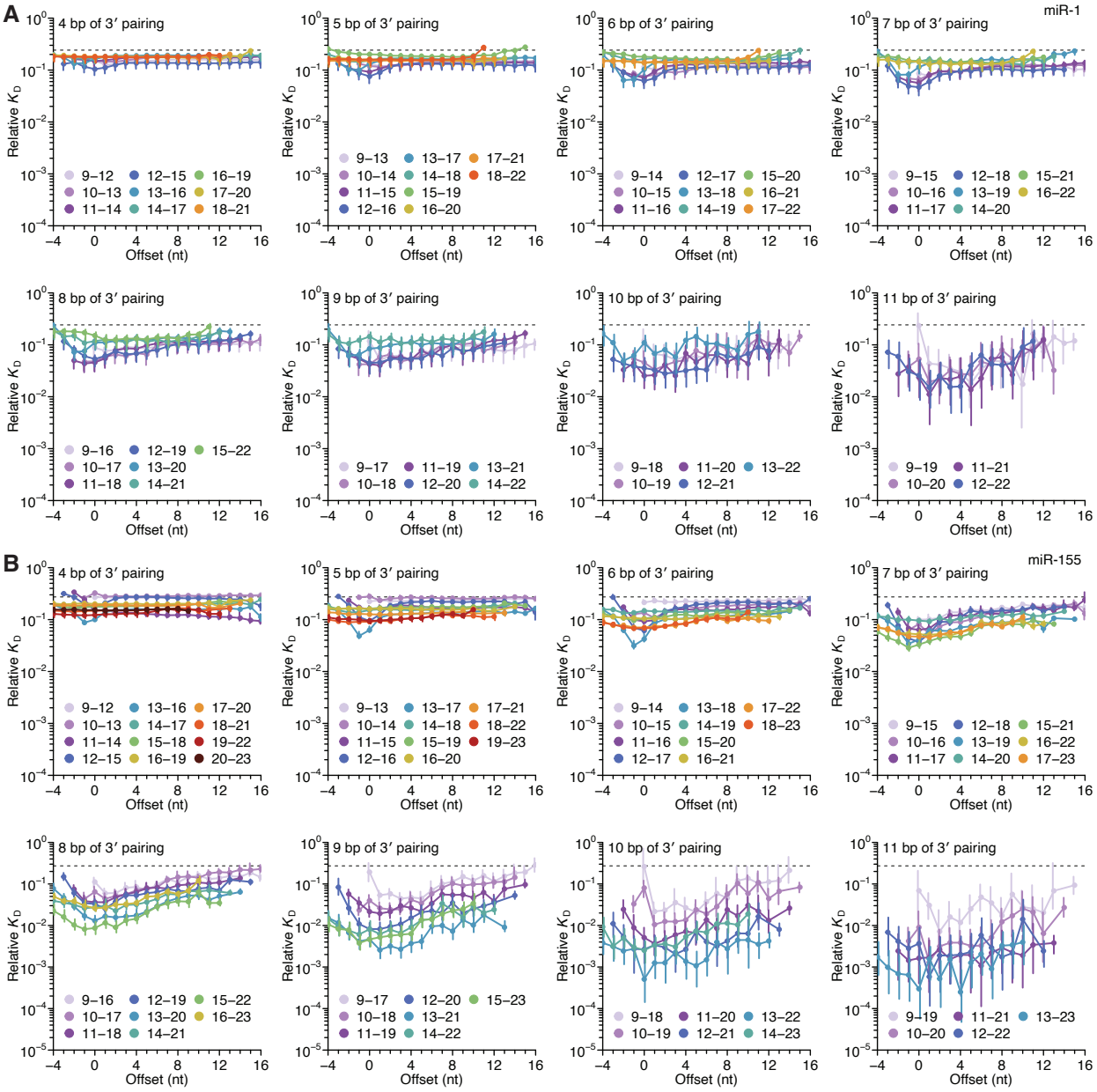
(**A**) Cumulative distributions of relative K_D values for let-7a 3′-compensatory sites that have 4 (orange) to 11 (dark blue) contiguous base pairs of 3′ pairing, from a replicate experiment depicted on the *y*-axis of both panels of Figure S1A. Everything is as in Figure 2A. (**B**) Relative K_D values of let-7a 3′-compensatory sites that had optimally positioned 3′-pairing of lengths 4–11 bp, from a replicate experiment depicted on the *y*-axis of both panels of Figure S1A. Otherwise, this panel is as in Figure 2B. (**C**) The dependency of canonical site binding affinity on library position, for let-7a (left), miR-1 (middle) and miR-155 (right). The dashed horizontal line indicates the geometric mean of the 18 relative K_D values of the seed mismatch sites, each calculated from reads with <4 nt of contiguous complementarity to the miRNA 3′ end. Library position is defined as illustrated in Figure 1C.



Supplemental Figure 3. Reanalysis of binding experiments performed in Becker, Ober-Reynolds et al. (2019).

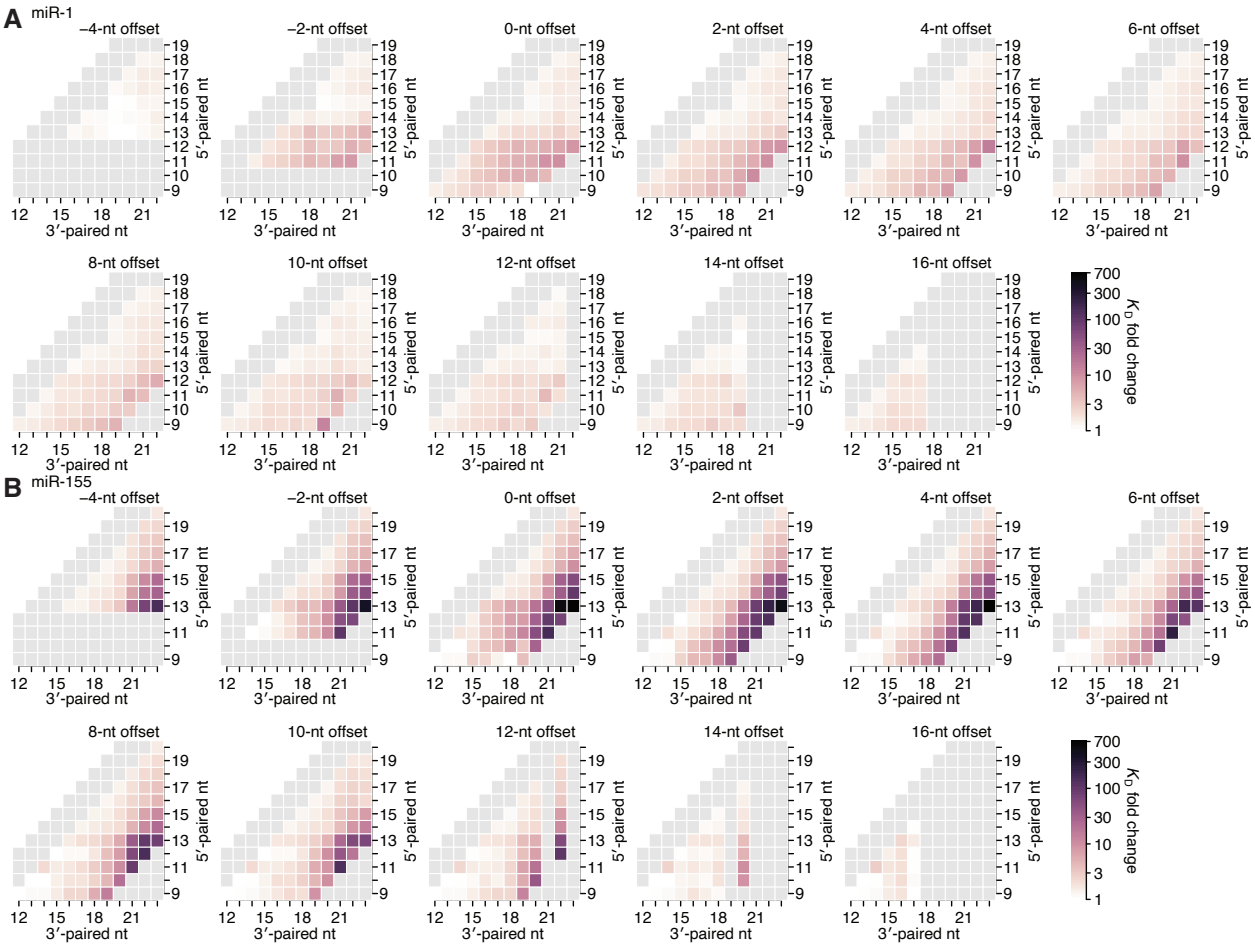
(**A**–**F**) Partial affinity profile of the let-7a 3′ region in the context of the 8mer (A), 7mer-m8 (B), 7mer-A1 (C), 6mer (D), 8mer-xC7 (E), and 8mer-xA5 (F) seed sites, using binding affinity measurements calculated using imaging-based, high-throughput single-molecule experiments (Becker et al., 2019). Each cell indicates the fold-change in K_D attributed to a 3′ site with indicated length, position, and offset of pairing. In (A–E), the fold-change is with respect to the geometric K_D of all the target RNAs with the corresponding seed site and <4 bp of 3′ pairing, and

in (F), the fold-change is with respect to 10 nM, because there were no target RNAs with a 8mer-xA5 site and no 3′ pairing. Gray boxes indicate pairing possibilities for which there was no data. Otherwise, this panel is as in Figure 2D.



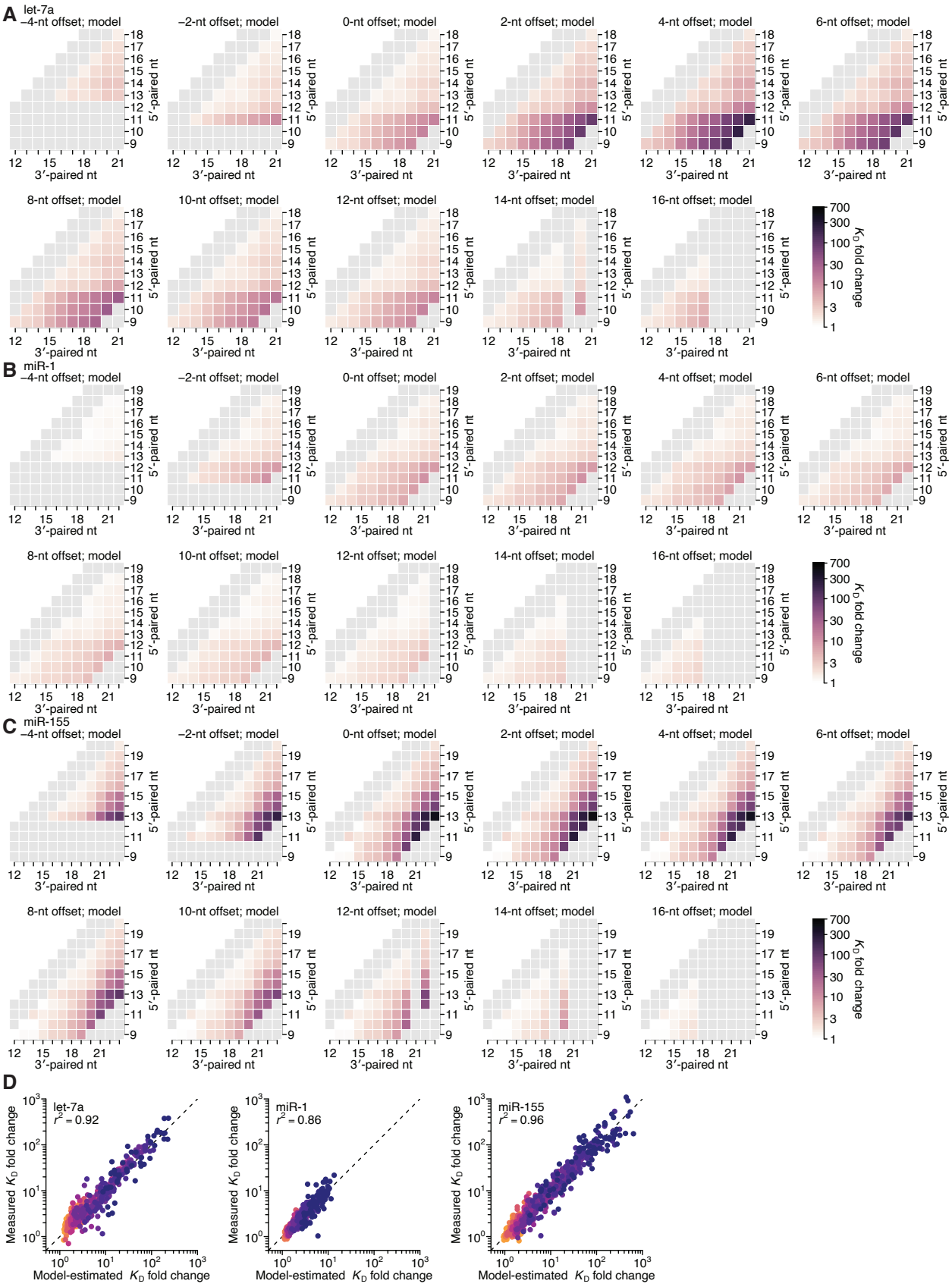
Supplemental Figure 4. Length, position, and offset trends of miR-1 and miR-155 indicate one binding mode.

(**A** and **B**) The dependency of miR-1 (A) and miR-155 (B) 3′-pairing on pairing length, position, and offset. Each panel shows the relative K_D values for 3′ of a specified length over a range of positions and offsets, spanning positions 9 (light violet) to 18 (red) when possible. Otherwise, this panel is as in Figure 2C.

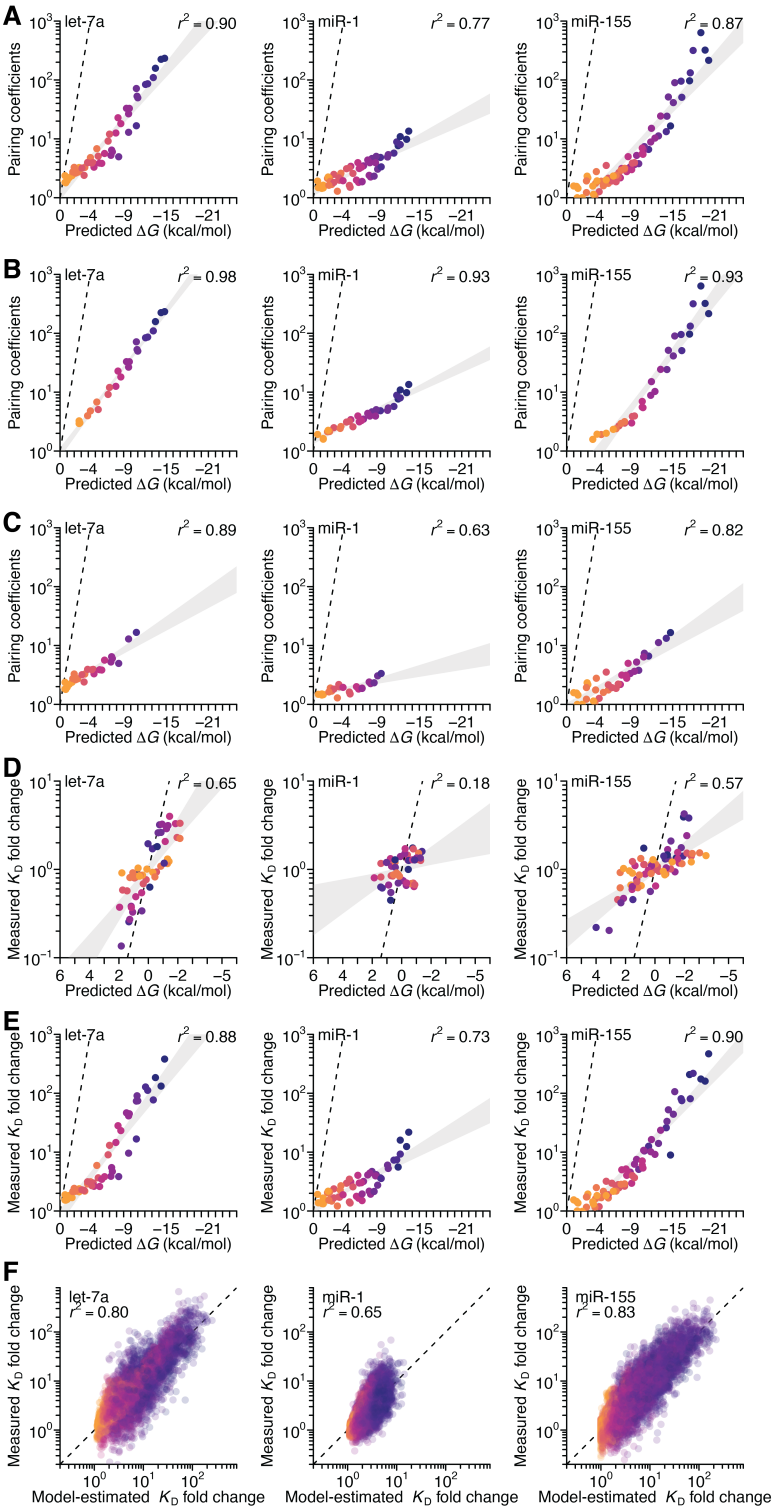


Supplemental Figure 5. Variation in miR-1 and miR-155 3′ pairing with different pairing lengths, positions, and offsets.

(**A** and **B**) Affinity profile of the miR-1 (A) and miR-155 (B) 3′ regions. Otherwise, this panel is as in Figure 2D..

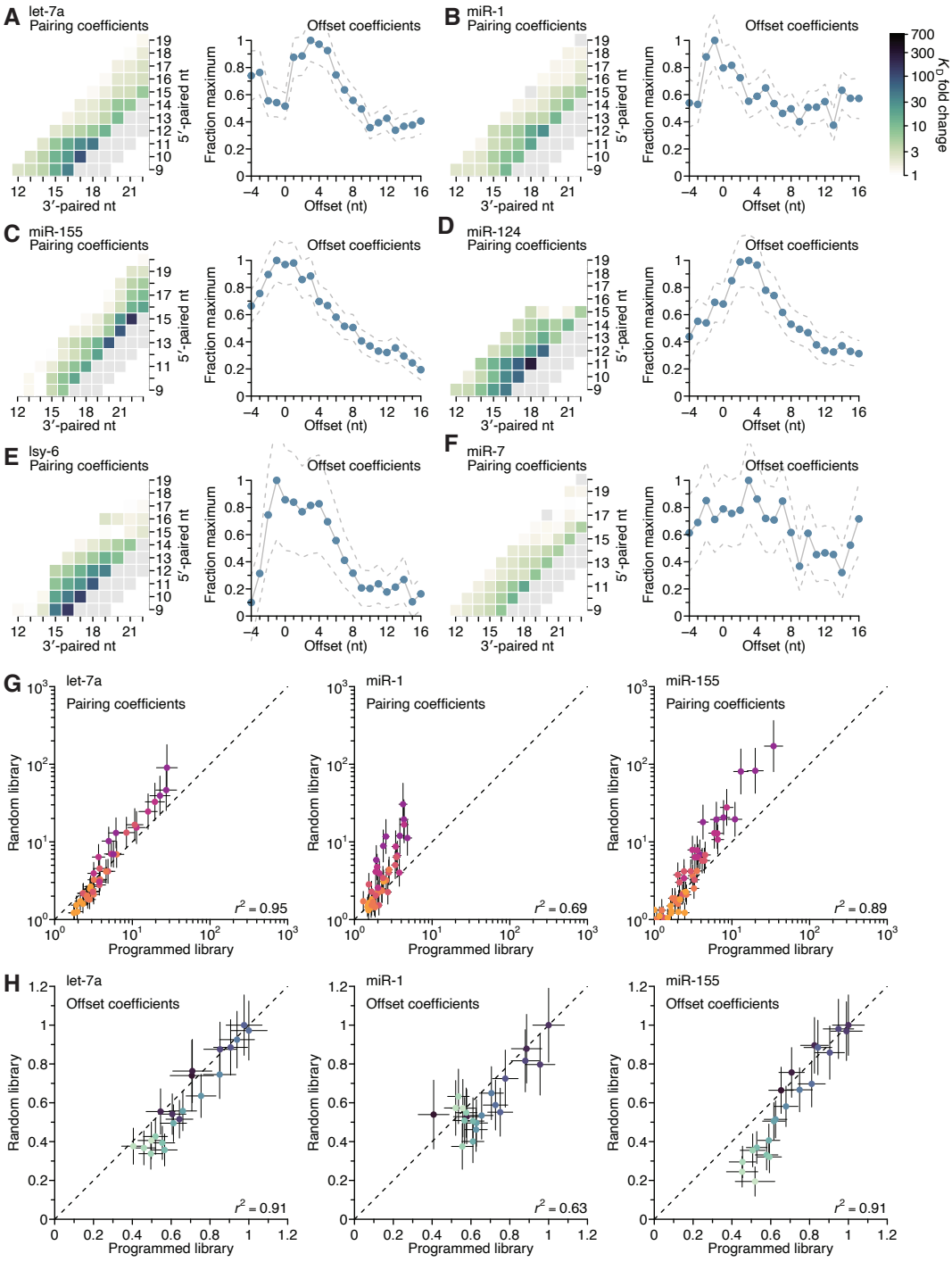

A let-7a
−4-nt offset; model
−2-nt offset; model
0-nt offset; model
2-nt offset; model
4-nt offset; model
6-nt offset; model
8-nt offset; model
10-nt offset; model
12-nt offset; model
14-nt offset; model
16-nt offset; model
5′-paired nt
3′-paired nt
K_D fold change
B miR-1
C miR-155
D
let-7a
r^2 = 0.92
miR-1
r^2 = 0.86
miR-155
r^2 = 0.96
Measured K_D fold change
Model-estimated K_D fold change

Supplemental Figure 6. Model prediction of 3′-compensatory pairing.
(**A**–**C**) Model-predicted affinity profiles of the let-7a (A), miR-1 (B), and miR-155 (C) 3′-regions. Gray cells are either those corresponding to pairing lengths <4 bp or >11 bp, or those for which there were no measured K_D fold change values for comparison (Figures 2D and S5), such as the let-7a 3′-compensatory sites ending with pairing to miRNA nucleotide 19 and with a 14-nt offset. Otherwise, this panel is as in Figure 2D. (**D**) Pairwise comparison of the model-predicted and measured K_D fold-change values for let-7a (left), miR-1 (middle), and miR-155 (right), with sites that have 4 (orange) to 11 (dark blue) contiguous bp of 3′ pairing. The r^2 reports on the coefficient of determination between the log-transformed predicted and measured values.



Supplemental Figure 7. Analysis of the correspondence of pairing coefficients with predicted Δ*G*, and performance of seed mismatch–effect model.
(**A**) The relationship between the model-derived pairing coefficients (Figure 4A–4C, left) and the predicted Δ*G* values (Figure 4D), when not controlling for length. Otherwise, this panel is as in Figure 4E. (**B** and **C**) The relationship between the model-derived pairing coefficients (Figure

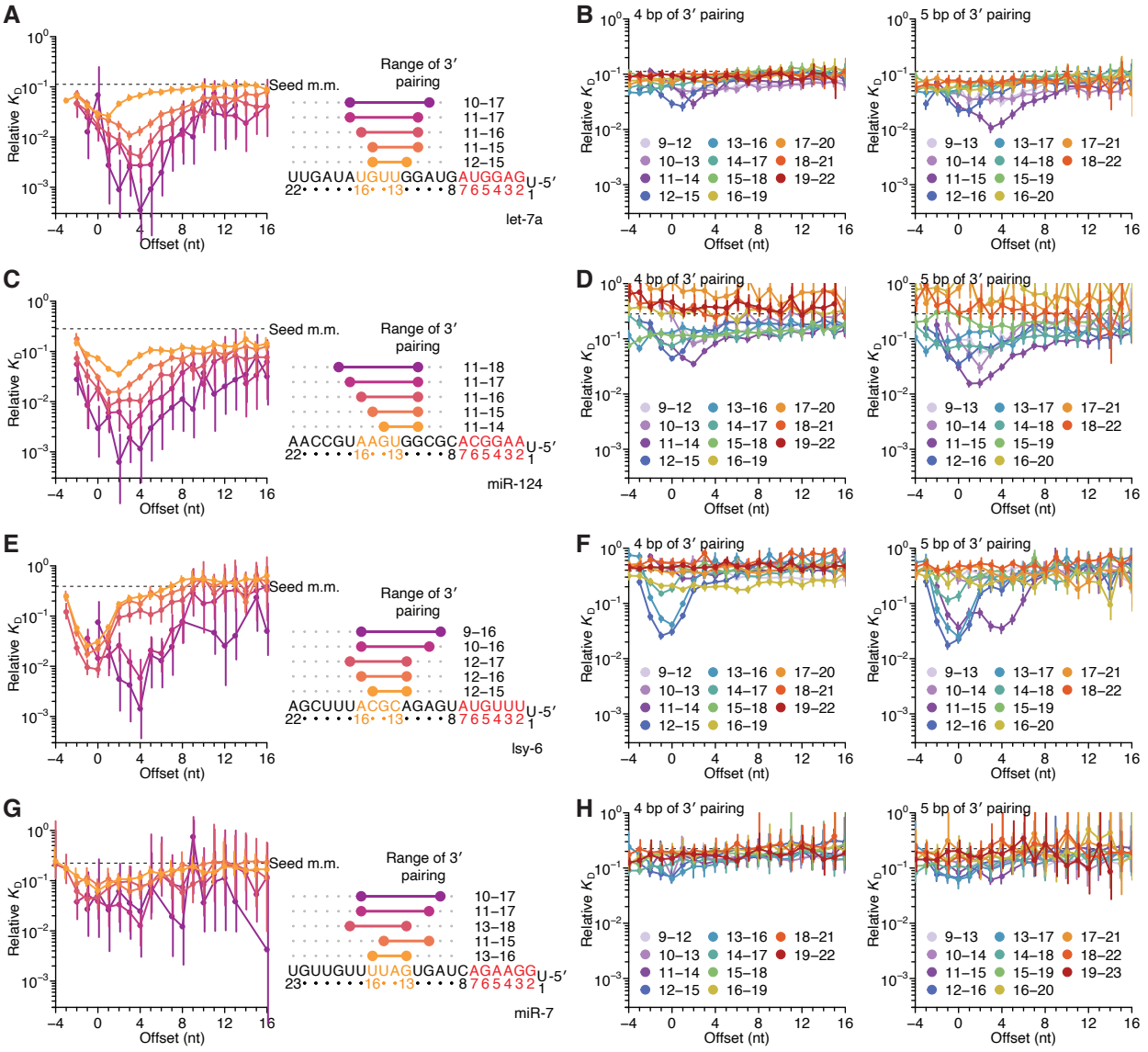
4A–4C, left) and the predicted Δ*G* values (Figure 4D), when separating the pairing based on whether it includes (B) or excludes (C) pairing to nucleotide 11 for let-7a (left), 12 for miR-1 (middle), and 20 for miR-155 (right). Otherwise, this panel is as in (A). (**D**) The relationship between the K_D fold-change values measured at their optimal offset and the predicted Δ*G* values (Figure 4D), when controlling for length. Otherwise, this panel is as in Figure 4E. (**E**) The relationship between the K_D fold-change values measured at their optimal offset and the predicted Δ*G* values (Figure 4D), when not controlling for length. Otherwise, this panel is as in (A). (**F**) Pairwise comparison of all of the model-predicted and measured K_D fold-change values for let-7a (left), miR-1 (middle), and miR-155 (right), when using an expanded model with pairing, offset, and seed-mismatch coefficients. Points are colored according the length of their pairing, which ranged from 4 (orange) to 11 (dark blue) contiguous bp. The r^2 reports on the coefficient of determination between the log-transformed predicted and measured values.



Supplemental Figure 8. Distinct pairing-range and offset preferences of different miRNAs in random-sequence AGO-RBNS experiments.

(**A**–**F**) Model-based analyses of 3′-pairing preferences of let-7a (A), miR-1 (B), miR-155 (C), miR-124 (D), lsy-6 (E), and miR-7 (F), for sites 4–8 bp in length, using data from previously reported, random-sequence AGO-RBNS experiments. Otherwise, these panels are the same as in 4A, left and middle-left. (**G**) Pairwise comparison of the pairing-range coefficients derived from the programmed library and the random libraries for let-7a (left), miR-1 (middle), and miR-155 (right), for 3′ pairing of lengths 4–11 bp. The r^2 reports on the coefficient of determination

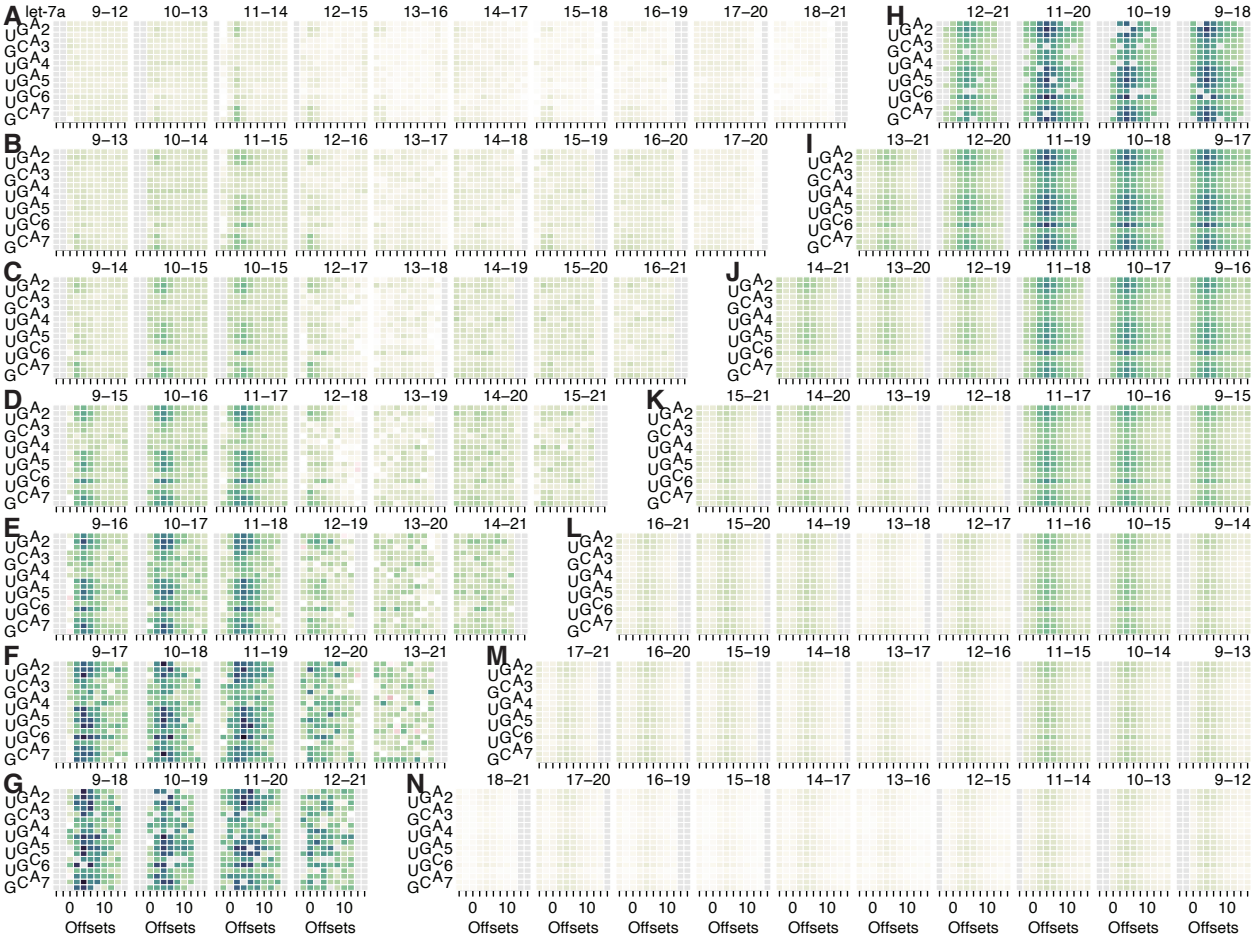
between the log-transformed values. (**H**) Pairwise comparison of the offset coefficients derived from the programmed and random libraries, for 3′ pairing of lengths 4–11 bp. The r^2 reports on the coefficient of determination between the log-transformed values.



Supplemental Figure 9. Identification of two binding modes for most miRNAs in random-library AGO-RBNS data.

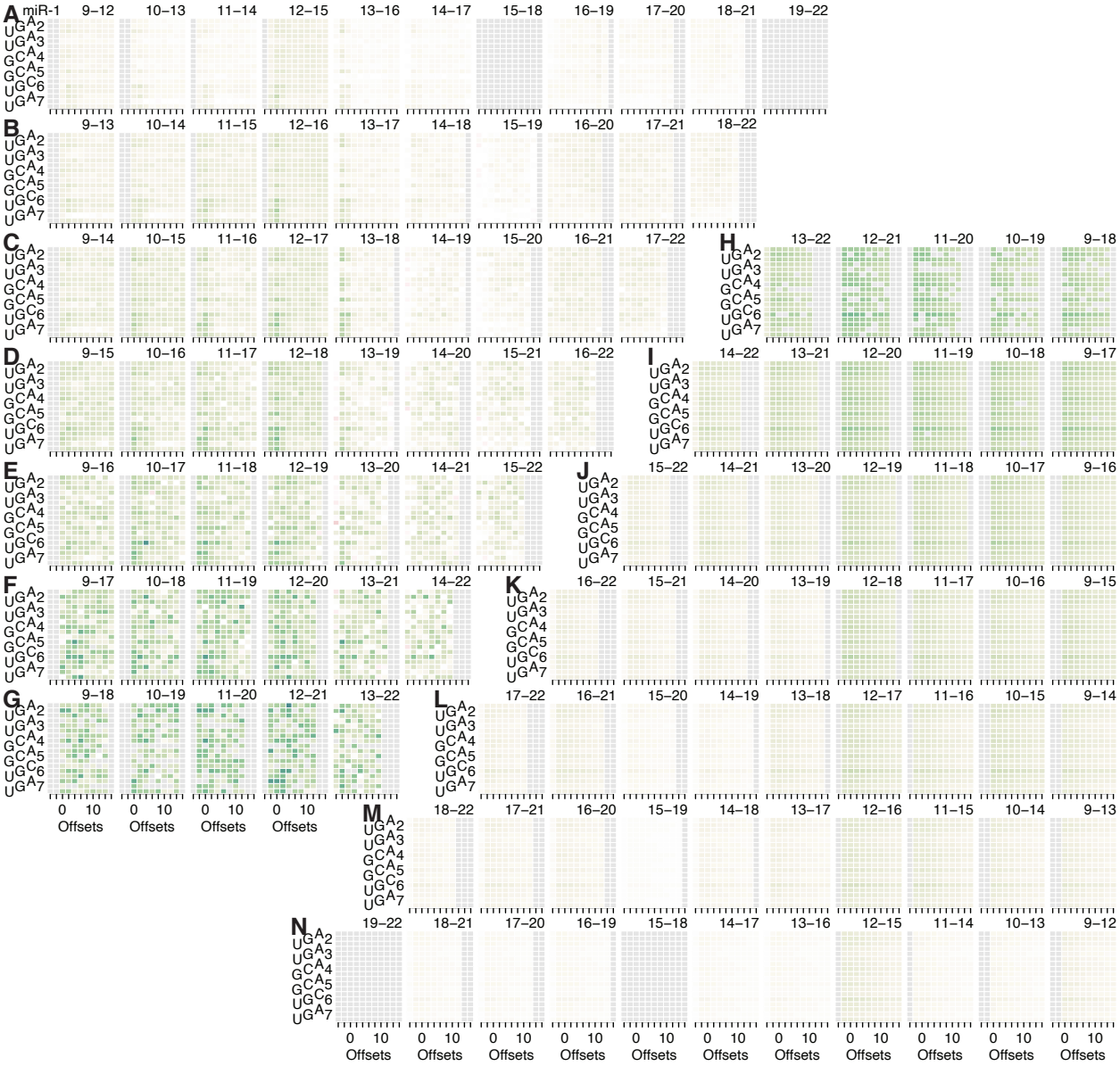
(**A**) Relative K_D values of let-7a 3′-compensatory sites that had optimally positioned 3′ pairing of lengths 4–8 bp, as measured with data obtained previously from fully randomized libraries (McGeary et al., 2019). Otherwise, this panel is as in Figure 2B. (**B**) The dependency of let-7a 3′ pairing affinity on pairing length, position, and offset, for 3′ pairing of 4 (left) and 5 (right) bp of pairing. Otherwise, this panel is as in Figure 2C.

(**C**–**H**) Equivalent analyses to those of (A) and (B), performed with miR-124 (C and D), lsy-6 (E and F), and miR-7 (G and H), as measured with data obtained previously from fully randomized libraries.



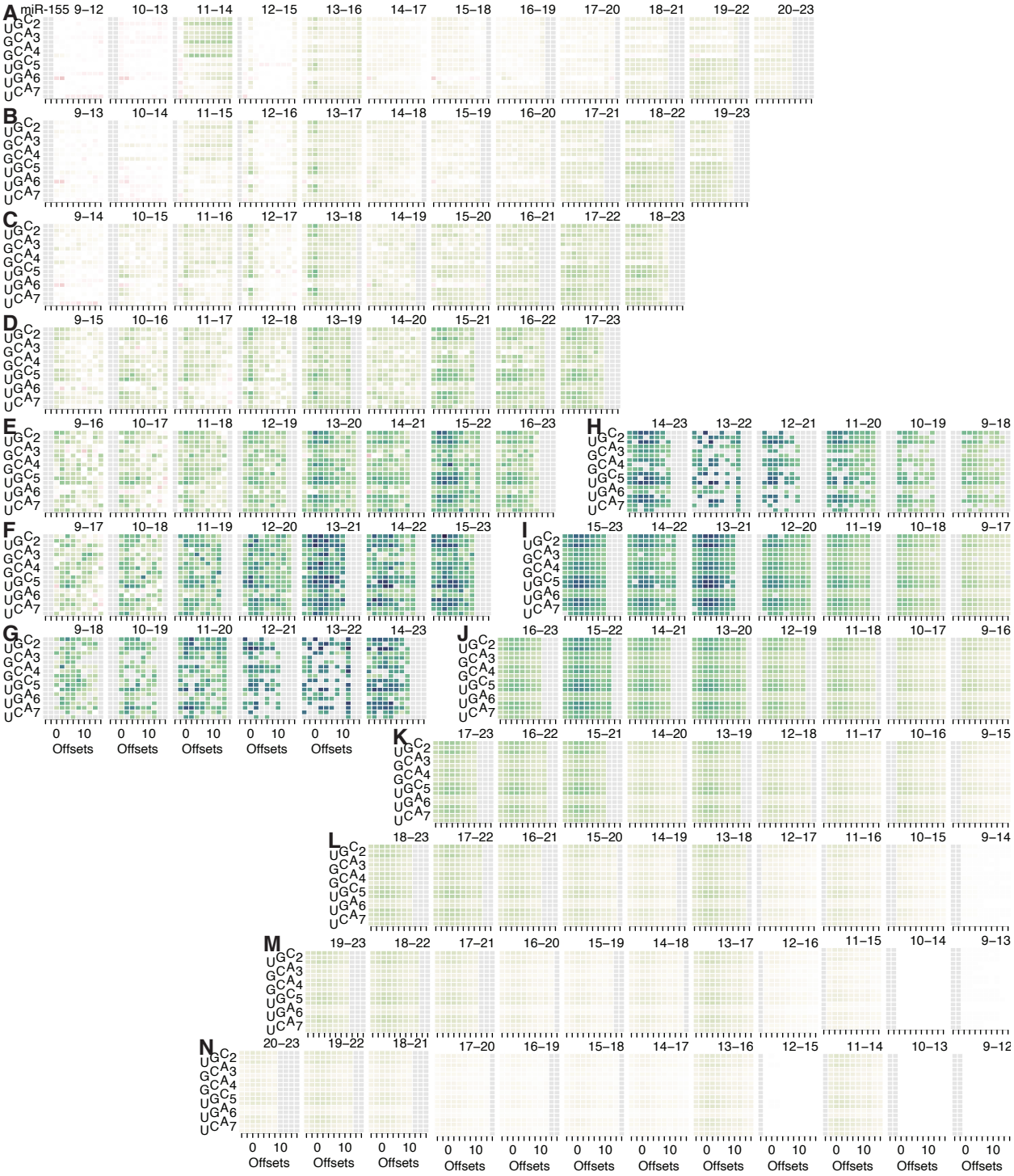
Supplemental Figure 10. Comparison of pairing-range, offset, and seed-mismatch model with measured K_D fold-change data for let-7a.

(**A**–**N**) Visual representation of the model performance evaluated in Figure S7F, left. The heat maps within the upper-left triangle (A–G) show the measured K_D values, while the heat maps within the lower-right triangle (H–N) show the model predictions. The two triangular arrays of heat maps are rotationally symmetric. Each individual heat map describes all of the variation associated with one defined stretch of 3′ pairing. Within each heat map, each row corresponds to a different seed mismatch type, with the left-hand numbers referring to the seed-mismatch position of every three rows, the staggered left-hand letters designating the mismatch identity of each row, and the columns each corresponding to a different offset of pairing.



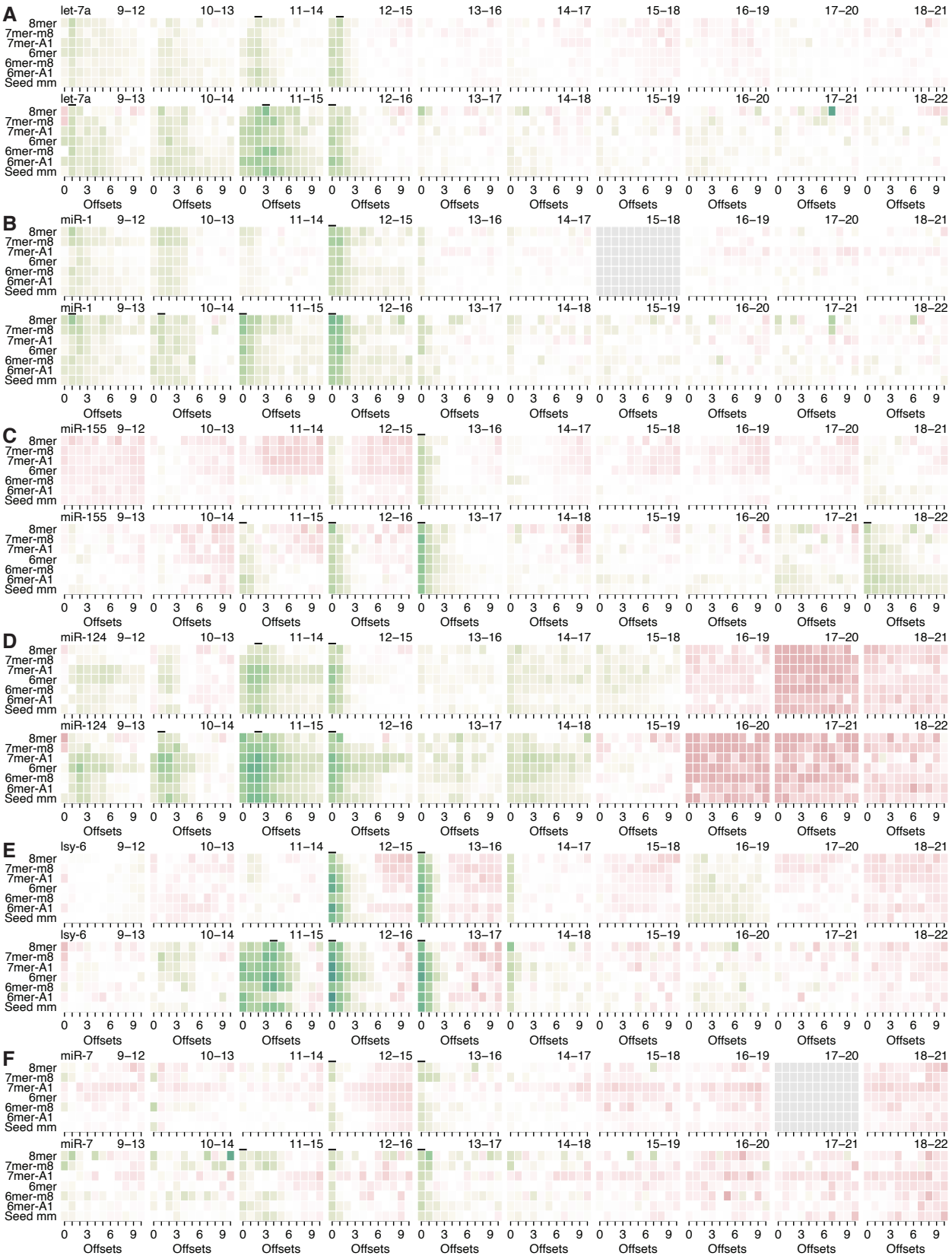
Supplemental Figure 11. Comparison of pairing-range, offset, and seed-mismatch model with measured K_D fold-change data for miR-1.

(**A**–**N**) Visual representation of the model performance evaluated in Figure S7F, middle. No model-predicted and measured K_D fold-change values are reported for pairing to positions 15–18 or 19–22 because they are the same sequence and thus unassignable. Otherwise, this figure is as in Figure S10.



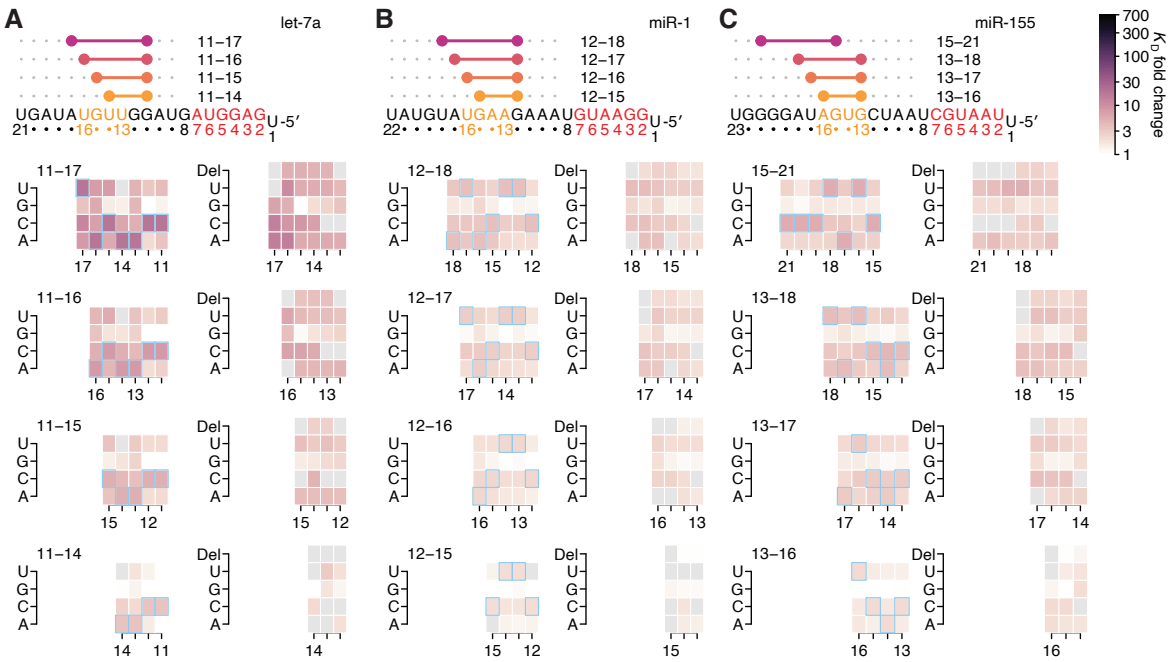
Supplemental Figure 12. Comparison of pairing-range, offset, and seed-mismatch model with measured K_D fold-change data for miR-155.

(**A**–**N**) Visual representation of the model performance evaluated in Figure S7F, right. Otherwise, this figure is as in Figure S10.


A
let-7a
9–12
10–13
11–14
12–15
13–16
14–17
15–18
16–19
17–20
18–21
let-7a
9–13
10–14
11–15
12–16
13–17
14–18
15–19
16–20
17–21
18–22
B
miR-1
C
miR-155
D
miR-124
E
lsy-6
F
miR-7
8mer
7mer-m8
7mer-A1
6mer
6mer-m8
6mer-A1
Seed mm
0 3 6 9
Offsets

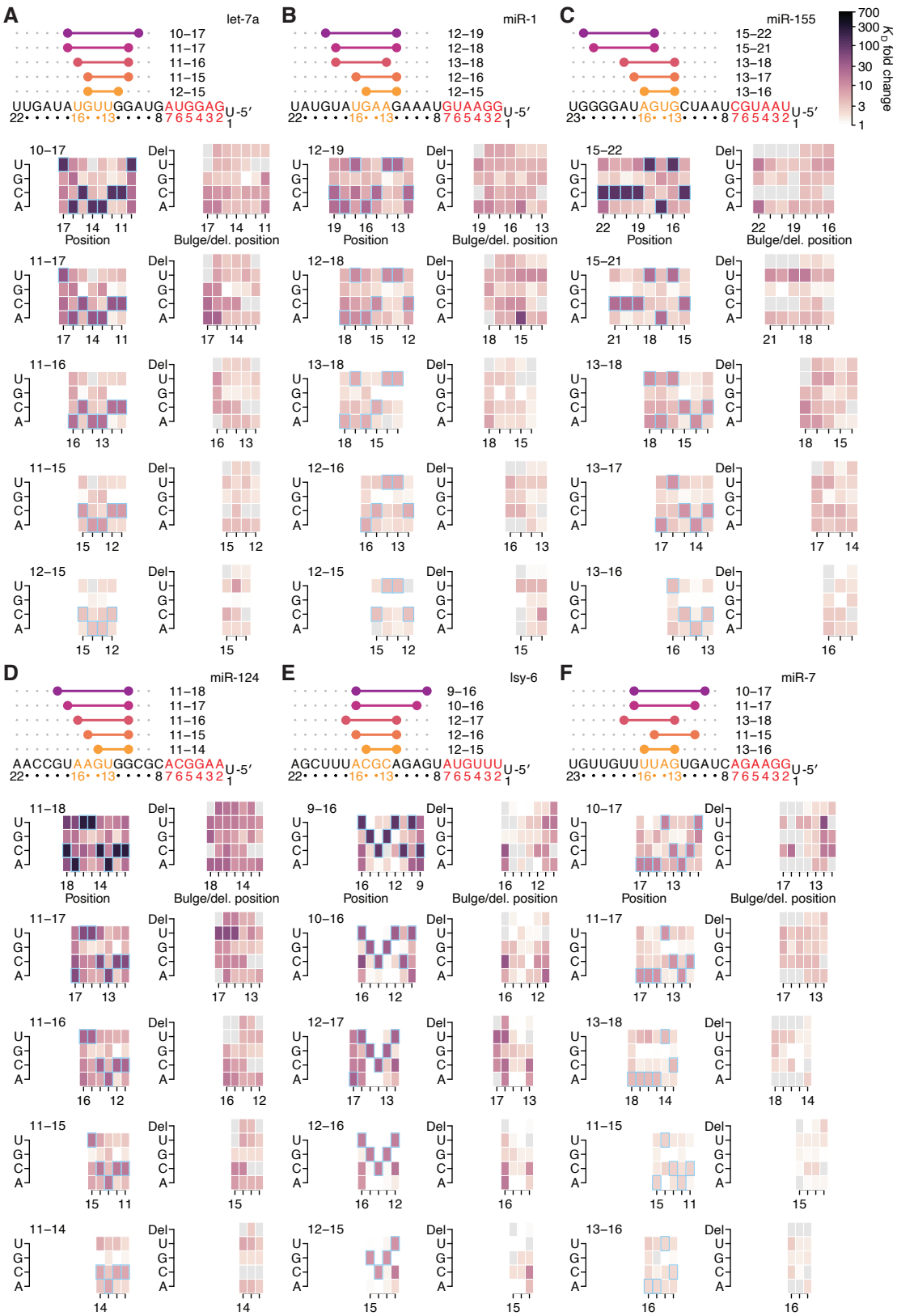
Supplemental Figure 13. Figure S13. Minimal influence of seed site type on 3′-supplementary pairing.

(**A**–**F**) Heat maps depicting the K_D fold change of canonical sites and seed mismatch sites of let-7a (A), miR-1 (B), miR-155 (C), miR-124 (D), lsy-6 (E), and miR-7 (F), with data obtained previously from fully randomized libraries. Each individual heat map describes one pairing range to the miRNA 3′ shown in the upper-right, with each row describing one of each canonical site or a seed mismatch site given by summing the read counts of all 18 internal 8mer-mismatch sites, and each column a different offset value ranging from 0 to +10 nt. Each cell is colored as in Figure 4A, left, with green and red indicating increased and decreased binding affinity, respectively. Those columns with horizontal lines above them were used to compute the average 3′-supplementary pairing for the six canonical sites used for the analysis in Figure 4H. K_D fold-change values were not calculated for pairing to positions 15–18 for miR-1 nor for pairing to positions 17–20 for miR-7 because these two segments were identical to the 4-nt segment at the very 3′ end of each respective miRNA sequence.



Supplemental Figure 14. Further analysis of the impact of mismatched, bulged, and deleted target nucleotides on 3′-compensatory pairing.

(**A**–**C**) The effect of mismatched, bulged, and deleted target nucleotides on 3′-compensatory pairing to let-7a (A), miR-1 (B), and miR-155 (C), in the context of the optimal sites 4–7 bp in length for each miRNA. Otherwise, these panels are as in Figure 7A.


A
let-7a
10–17
11–17
11–16
11–15
12–15
UUGAUAUGUUGGAUGAUGGAG
22 16 13 8 7 6 5 4 3 2 U-5′ 1
B
miR-1
12–19
12–18
13–18
12–16
12–15
UAUGUAUGAAGAAAUGUAAGG
22 16 13 8 7 6 5 4 3 2 U-5′ 1
C
miR-155
15–22
15–21
13–18
13–17
13–16
UGGGGAUAGUGCUAAUCGUAAU
23 16 13 8 7 6 5 4 3 2 U-5′ 1
700
300
100
30
10
3
1
K_D fold change
D
miR-124
11–18
11–17
11–16
11–15
11–14
AACCGUAAGUGGCGCACGGAA
22 16 13 8 7 6 5 4 3 2 U-5′ 1
E
lsy-6
9–16
10–16
12–17
12–16
12–15
AGCUUUACGCAGAGUAUGUUU
22 16 13 8 7 6 5 4 3 2 U-5′ 1
F
miR-7
10–17
11–17
13–18
11–15
13–16
UGUUGUUUUAGUGAUCAGAAGG
23 16 13 8 7 6 5 4 3 2 U-5′ 1
Del
U
G
C
A
Position
Bulge/del. position

Supplemental Figure 15. The impact of mismatched, bulged, and deleted target nucleotides on 3′-compensatory pairing from random-sequence AGO-RBNS experiments.
(**A**–**F**) The effect of mismatched, bulged, and deleted target nucleotides on 3′-compensatory pairing to let-7a (A), miR-1 (B), and miR-155 (C), miR-124 (D), lsy-6 (E), and miR-7 (F), in the context of the optimal sites 4–8 bp in length for each miRNA, with data obtained previously from fully randomized libraries. Otherwise, these panels are as in Figure 7A.

References

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, e05005.

Alper, J.S., and Gelb, R.I. (1990). Standard errors and confidence intervals in nonlinear regression: comparison of Monte Carlo and parametric statistics. J. Phys. Chem. *94*, 4747–4751.

Ameres, S.L., Horwich, M.D., Hung, J.-H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P.D. (2010). Target RNA–directed trimming and tailing of small silencing RNAs. Science *328*, 1534–1539.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bartel, D.P. (2018). Metazoan microRNAs. Cell *173*, 20–51.

Becker, W.R., Ober-Reynolds, B., Jouravleva, K., Jolly, S.M., Zamore, P.D., and Greenleaf, W.J. (2019). High-throughput analysis reveals rules for target RNA binding and cleavage by AGO2. Mol. Cell *75*, 741–755.e11.

Bitetti, A., Mallory, A.C., Golini, E., Carrieri, C., Gutiérrez, H.C., Perlas, E., Pérez-Rico, Y.A., Tocchini-Valentini, G.P., Enright, A.J., Norton, W.H.J., et al. (2018). MicroRNA degradation by a conserved target RNA regulates animal behavior. Nat. Struct. Mol. Biol. *25*, 244–251.

Brancati, G., and Großhans, H. (2018). An interplay of miRNA abundance and target site architecture determines miRNA activity and specificity. Nucleic Acids Res. *46*, gky201-.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA–target recognition. PLOS Biol. *3*, e85.

Cazalla, D., Yario, T., Steitz, J.A., and Steitz, J. (2010). Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. Science *328*, 1563–1566.

Chen, G.R., Sive, H., and Bartel, D.P. (2017). A seed mismatch enhances Argonaute2-catalyzed cleavage and partially rescues severely impaired cleavage found in fish. Mol. Cell *68*, 1095–1107.e5.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes Dev. *18*, 504–511.

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Nostrand, E.L.V., Pratt, G.A., et al. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Mol. Cell *70*, 854–867.e9.

Ecsedi, M., Rausch, M., and Großhans, H. (2015). The let-7 microRNA directs vulval development through a single target. Dev. Cell *32*, 335–344.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol. *5*, R1.

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

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

Han, J., LaVigne, C.A., Jones, B.T., Zhang, H., Gillett, F., and Mendell, J.T. (2020). A ubiquitin ligase mediates target-directed microRNA decay independently of tailing and trimming. Science eabc9546.

Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nat. Rev. Genet. *16*, 421–433.

Kleaveland, B., Shi, C.Y., Stefano, J., and Bartel, D.P. (2018). A network of noncoding regulatory RNAs acts in the mammalian brain. Cell *174*, 350–362.e17.

Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., Piedade, I. da, Gunsalus, K.C., Stoffel, M., et al. (2005). Combinatorial microRNA target predictions. Nat. Genet. *37*, 495–500.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–900.

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

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

Lorenz, R., Bernhart, S.H., Siederdissen, C.H. zu, Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Mata, M. de la, Gaidatzis, D., Vitanescu, M., Stadler, M.B., Wentzel, C., Scheiffele, P., Filipowicz, W., and Großhans, H. (2015). Potent degradation of neuronal miRNAs induced by highly complementary targets. EMBO Rep. *16*, 500–511.

McGeary, S.E., Lin, K.S., Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., and Bartel, D.P. (2019). The biochemical basis of microRNA targeting efficacy. Science *366*, eaav1741.

Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Dev. Biol. *267*, 529–535.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901–906.

Sætrom, P., Heale, B.S.E., Snøve, O., Aagaard, L., Alluin, J., and Rossi, J.J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res. *35*, 2333–2342.

Salomon, W.E., Jolly, S.M., Moore, M.J., Zamore, P.D., and Serebrov, V. (2015). Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. Cell *162*, 84–95.

Schirle, N.T., Sheu-Gruttadauria, J., and MacRae, I.J. (2014). Structural basis for microRNA targeting. Science *346*, 608–613.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Sheu-Gruttadauria, J., Xiao, Y., Gebert, L.F., and MacRae, I.J. (2019a). Beyond the seed: structural basis for supplementary microRNA targeting by human Argonaute2. EMBO J. *38*, e101153.

Sheu-Gruttadauria, J., Pawlica, P., Klum, S.M., Wang, S., Yario, T.A., Oakdale, N.T.S., Steitz, J.A., and MacRae, I.J. (2019b). Structural basis for target-directed microRNA degradation. Mol. Cell *75*, 1243–1255.e7.

Shi, C.Y., Kingston, E.R., Kleaveland, B., Lin, D.H., Stubna, M.W., and Bartel, D.P. (2020). The ZSWIM8 ubiquitin ligase mediates target-directed microRNA degradation. Science eabc9359.

Tomari, Y., and Zamore, P.D. (2005). Perspective: machines for RNAi. Genes Dev. *19*, 517–529.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. Nature *505*, 706–709.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. Cell *151*, 1055–1067.

Chapter 4. Future directions

Further advances in miRNA targeting

The work described in Chapter 2 advances our understanding of animal miRNA targeting, through the application of high-throughput biochemistry to learn vast numbers of relative K_D values corresponding to miRNA binding sites throughout the transcriptome. This advancement was driven in part by the miRNA-specific nature of the predictions, which had not been possible due the challenges of learning miRNA-specific features solely from in vivo data, and partly by the use of a biochemically informed model rather than multiple linear regression (Agarwal et al., 2015). This framework, when correcting the median r^2 values for measurement noise, is able to capture ~60% of the effects of a miRNA when using empirically measured relative K_D values, and ~50% when using CNN-predicted relative K_D values. The 10% loss of prediction when using CNN-predicted values is likely due to an insufficient quantity of data with which to train the CNN. To this end, performing AGO-RBNS with a greater diversity of miRNA sequences, performing more miRNA transfection experiments, and possibly modifying the architecture of the CNN upon acquisition of these data, would be expected to cause the CNN-derived predictions to approach the performance achieved for individual miRNAs for which measured affinity values exist.

Improvement of prediction beyond that of 60% when using "true" K_D values (i.e., when either empirically measured, or derived from an asymptotically perfect CNN) will be less straightforward, and will require thoughtful consideration of what remains unknown about the miRNA pathway and regulated mRNA metabolism more generally. The equation used in Chapter 2 for predicting repression of an mRNA from the binding affinities of its sites gave the best performance among those tested, and while it accurately models site occupancy according to

biochemical principles, the repressive action of the miRNA performed on the mRNA is collapsed into a single value, the parameter *b* (equation 2.53). Indeed, that a single *b* value performs as well as it does is consistent with a model in which the function of a miRNA, once bound, is largely consistent between target sites and between different mRNA sequences. This notion is echoed by reports showing that the amount of repression experienced by a target mRNA is minimally influenced by the basal decay rate of that mRNA (Eisen et al., 2020; Larsson et al., 2010), suggesting miRNAs increase the deadenylation rate of their targets by a constant factor, rather than adding to the deadenylation rate through a dinstinct molecular mechanism. However, the current lack of quantitative details regarding the complex nature of the association of Ago–miRNA complexes with GW182/TNRC6 proteins, PAN2–PAN3, and CCR4NOT, and how this relates to deadenylation rate, prevents identification of any molecular circumstances in which this simplistic model of repression is ill-suited (Braun et al., 2011, 2012; Eulalio et al., 2008).

It is becoming increasingly appreciated that many biological processes occur in biomolecular condensates, which physically form a distinct phase than that of the surrounding cytoplasm or nucleus (Banani et al., 2017; Hnisz et al., 2017). Indeed, many of the components of the basal mRNA destabilization machinery, as well as miRNAs, are preferentially enriched in cytoplasmic granules known as P bodies (Luo et al., 2018; Parker and Sheth, 2007). Because the cellular relationship between P-body enrichment of an mRNA and its capacity to undergo miRNA-mediated repression are not well understood, all miRNA target–prediction models are forced to make the assumption that both basal degradation and miRNA-specific degradation occur equally throughout the cytoplasm. Experiments measuring the basal P-body enrichment of mRNAs, and the change in both their P-body enrichment and abundance in response to miRNA

induction, will be required to probe the importance of considering P-body localization in models of miRNA targeting more generally.

The discovery that a phosphorylation cycle acting on AGO is necessary for efficient miRNA-mediated repression (Golden et al., 2017; Huberdeau et al., 2017) is particularly intriguing, partially because such a cycle challenges a simplistic model in which miRNAs repress their targets through a constitutive activity of GW182/TNRC6 recruitment, with target occupancies determined by intrinsic association and dissociation rates. In particular, the finding that AGO is phosphorylated upon target binding, but that the phosphorylated AGO population is specifically unable to associate with target RNAs, appears superficially paradoxical (Golden et al., 2017), and could be interpreted to imply that AGO–miRNA complexes participate in many short-lived interactions, even with high-affinity sites, due to target ejection upon AGO phosphorylation. In vivo measurements of the characteristic length of time between target binding, AGO phosphorylation, and target unbinding would be informative for clarifying the role of the phosphorylation cycle, as would in vitro reconstitution of the system to define the minimal requirements for phosphorylation and its acute effect on the AGO–miRNA–target RNA ternary complex. A speculative model for how the phosphorylation cycle facilitates miRNA-mediated repression is that it facilitates trafficking of individual mRNAs to P bodies[1], by causing *increased* association of the miRNA with the target RNA. In this model, the phosphorylated AGO is unable to bind other targets, or to be purified via the capture-competitor method (Flores-Jasso et al., 2013), because it is still engaged with the target that promoted its phosphorylation. The increased dwell time of phosphorylated AGO with its bound target sites could rapidly promote higher-order AGO–miRNA–target RNA–TNRC6 assemblies, thus favoring their

[1]Or, more generically, to any localized sites of mRNA decay within the cytoplasm.

inclusion in P bodies, which are enriched for GW182/TNRC6 proteins. Dephosphorylation of AGO once inside the P body would enable target release, and enable eventual exit of both the AGO–miRNA complex and the target RNA from the granule. This hypothesis could be tested by querying for differential P-body enrichment of miRNAs and top target RNAs upon perturbations of either the ANKRD52–PPP6C phosphatase complex or the CSNK1A1 kinase, as well as through live-cell imaging of fluorescently tagged CSNK1A1, ANKRD52–PPP6C, and wild-type and kinase-resistant AGO.

Another avenue by which miRNA targeting prediction could be improved is through improved prediction of in vivo site accessibility. In the past decade, a number of techniques have been developed for detection of RNA structural accessibility, such as SHAPE-seq (Lucks et al., 2011), DMS-seq (Rouskin et al., 2014), and DMS-MaPseq (Zubradt et al., 2017), which all rely on some variation of non-sequence-specific chemical modification of RNA coupled to either early truncation of cDNA due to the inability of the reverse transcriptase to process past modified nucleotides, or nucleotide conversion using a specialized polymerase. One major obstacle in applying these approaches for improved prediction of miRNA targeting more generally lies in the vast number of reads required to accurately query the structural information of any stretch of RNA: when considering a cutoff of 20-fold read coverage (a threshold at which the majority of mRNA sequence exhibits a Pearson $r > 0.9$ between replicates), fewer than 5% of expressed genes in HEK293T can be confidently analyzed from 100 million reads (Zubradt et al., 2017). Fortunately, this presents only a practical, rather than a theoretical, limitation for using such approaches. Indeed, with release of the Illumina NovaSeq, which generates upwards of 40 billion reads per run, it seems plausible that a greater number of high-, medium-, and low-abundance mRNAs will be increasingly structurally profiled, such that target predictions for a

greater number of 3′-UTR sequences could use measured structure maps rather than predictions from current RNA folding algorithms.

The accessibility of any RNA will also be influenced by the cohort of RBPs that bind the mRNA sequence. Accurate incorporation of other RBPs into quantitative models of miRNA target prediction is complex, as it should lead to reduced site accessibility in cis due to direct competition with overlapping miRNA and RBP binding sites, but increased accessibility in trans through partial disruption of secondary structural element that occlude other miRNA sites. To this end, the binding preferences of a large number of RBPs has been profiled (Dominguez et al., 2018), which in theory could enable their incorporation into miRNA models, but would require additional metadata such as 1) reference K_D values by which to scale the binding profiles for each RBP, 2) knowledge of the expression level of each RBP in comparison to that of the miRNAs, and 3) the downstream effects of each RBP on its transcriptomic targets once bound. To this end, a recent large-scale study reporting measurements of in vitro binding, in vivo binding, mRNA expression changes upon knockdown, chromatin association, and subcellular localization for hundreds of RBPs (Nostrand et al., 2020) might serve as a rich resource for the construction of parameters relevant to such integrated modeling of miRNA repression, RBP-based regulation, and the influence of RNA structure on both.

Indeed, consideration of incorporating both RNA secondary structure and the binding and regulatory properties of other expressed RBPs into models of miRNA-mediated repression motivates the broader question of how to identify and account for inaccuracies of prediction due to factors extrinsic to the miRNA pathway more generally. The secondary effects of miRNAs, caused by repression of primary targets that also participate in gene-regulatory roles, have not been extensively studied, due to the lack of straightforward ways by which to do so. Some

success has been had in deconvoluting the effects of miRNA–transcription factor networks through analysis of RNA-seq, AGO-iCLIP, and histone modification marks in both wild-type and Dicer knock-out immortalized murine fibroblasts (Gosline et al., 2016). In particular, the study identified transcriptional changes for many mRNAs by comparison of intronic and exonic reads, and further explained a statistically significant fraction of these mRNA-transcriptional changes through direct targeting of associated TFs by expressed miRNAs. More recently, a study distinguished primary from secondary targets though analysis of RNA-seq and Precision Run-On sequencing (PRO-seq) performed in parallel in HEK293T, as changes in transcription identified by PRO-seq could be used to assign false-positive changes observed in the RNA-seq (Patel et al., 2020). These approaches show promise for further improvement of miRNA prediction, perhaps by expanded intronic read–based analyses of nuclear-enriched mRNA, the acquisition of more nascent-RNA profiling datasets, or construction of quantitative maps relating expression changes in known TFs to expected gene-regulatory effects within the transcriptome.

An alternative, complementary approach that might aid in improved prediction of miRNA-mediated repression would be to develop a method that quantitatively reports on the occupancy of individual miRNA target sites in cells. In principle, CLIP could serve this purpose, if the sequence-specific crosslinking biases observed with this technique could be either experimentally eliminated or accurately corrected for. Indeed, recent development of time-resolved RNA–protein CLIP (kinetic CLIP, or KIN-CLIP), which uses a pulsed femtosecond UV laser, shows promise for precise measurement of association and dissociation kinetics (and thus occupancies) for individual RBPs with each of their target sites in vivo (Sharma et al., 2020). Performing miRNA transfection in mammalian cell culture, followed by KIN-CLIP and RNA-seq in parallel would provide insight into whether those mRNAs for which the predicted

repression is least accurate are explained by target-site occupancies in disagreement with that predicted using in vitro–measured or CNN-derived K_D values.

In vivo occupancies might also be learned through proximity labeling–based approaches. It has recently been shown that expression of cytosolic, ER-tethered, and nuclear APEX2 in HEK293T cells enabled specific isolation of differentially localized RNAs, by pre-incubation of the cells with biotin-tyramide and hydrogen peroxide, followed by incubation of total RNA with streptavidin-coated magnetic beads (Padrón et al., 2019). Such approaches could readily be adapted for studying miRNA targeting, by generating APEX2–AGO fusion proteins, and performing the pre-incubation reactions for varying amounts of time, and performing mild RNase digestion or RNA fragmentation while the RNA is immobilized on beads. Indeed, the occupancies inferred from the APEX-based approach could be compared to those inferred from KIN-CLIP, and both could be evaluated for their agreement with observed repression. One could further imagine performing "split APEX" (sAPEX) (Han et al., 2019), in which one half of the APEX protein is fused to AGO, and the other half incorporated adjacent to one of the AGO-binding hotspots within a TNRC6 protein (Elkayam et al., 2017; Pfaff et al., 2013). Comparison of the results of the sAPEX approach with those of the standard APEX approach could potentially confirm the existence of sites that are seemingly well bound by AGO but do not elicit repression, thereby enabling further study of the molecular bases of such discrepancies.

In evaluating the current state of miRNA targeting prediction, it is reasonable to consider what motivates its continued improvement—after all, the integrated transcriptomic response to a miRNA can be directly measured using RNA-seq, thus obviating the need for a target-prediction algorithm to describe the effects of a miRNA in any particular system. A superficial response to this is that improved models of target prediction will be useful for evaluating the role of miRNAs

in biological contexts for which such data are not easily generated, such as within particular cell types that are challenging to isolate from heterogenous tissue, or when inferring aberrant regulation from human genomic sequences. In such circumstances, the predictions themselves provide direct benefit to the biological research community.

An entirely distinct motivation for the advancement of miRNA targeting is its utility towards the larger goal of a complete, quantitative understanding of gene expression control. This is facilitated by both the unambiguously repressive effect of the pathway on its targets, and its reprogrammable specificity through the use of different guide sequences. As the predicted effects of miRNAs of different sequences increasingly improve, the individual inaccuracies of the predictions will enable insight into other regulatory processes. For example, any 3′-UTR sequence segments that are consistently under- or over-repressed compared to model predictions might constitute an example of unusual local structure or adjacency to RBPs with synergistic effector functions. Additionally, identification of persistent covariance among pairs of mRNAs across many miRNA-transfection experiments might facilitate further construction of gene-regulatory networks, thereby connecting observed patterns of regulation in the nucleus to those in the cytoplasm. Mechanistic investigation of a sufficient number of such examples might enable further quantitative insights that could be incorporated into predicted repression. From this perspective, a model of miRNA-mediated repression might be thought of as an incipient model of the total regulatory control of animal gene expression, and perhaps eventually, the total molecular biology of an animal cell.

Towards a quantitative definition of gene function

Constructing a model of a large-scale biological system is not a novel idea, as both a whole-cell model of the human urogenital parasite *Mycoplasma genitalium* (Karr et al., 2012), and a "half-cell" model of *E. coli* (Macklin et al., 2020) have been reported, which mathematically formalize and integrate the processes of growth, replication, metabolism, and RNA and protein production for both organisms. Such work, while not enabling discovery per se, provides essential insights regarding the consistency of disparate experimental results with one another, as it allows the parameter values from biological processes to functionally cross-talk with those of another within the model. For example, the *E. coli* model exhibited an immediate discrepancy between its simulated (125 min) and expected (44 min) doubling time, and was traced to insufficient abundance of RNA polymerase and ribosomes (Macklin et al., 2020). When adjusting the production rates of these molecules such that the doubling time agreed with measurement, it brought the RNA polymerase and ribosome abundances into line with more recently published measurements (from which the parameter values were not initially derived), and further enabled the model to recapitulate experimentally measured RNA mass, rRNA initiation rate, and ribosome elongation rate over the course of cell doubling time.

An integrated model of animal gene-regulatory processes in the cytoplasm would enable assessment of how well the growing number of published compendia of RBP binding preferences (Dominguez et al., 2018; Nostrand et al., 2020; Ray et al., 2013) and RBP activity (Nostrand et al., 2020) can be used to predict the absolute stability and translation rate of each mRNA. Such models could additionally incorporate quantitative data describing the influence of codon optimality (Presnyak et al., 2015), known sites of mRNA base modification such as pseudouridine (Carlile et al., 2014) and N^6-methyladenosine (Liu et al., 2019) and their

associated “reader” proteins that bind the modification and regulate RNA stability (Zaccara et al., 2019), and any other regulatory mechanisms for which large-scale quantitative data exist.

In the course of the experiments and analyses described in this thesis, quantitative binding maps were constructed for six miRNAs, which in principle serve as a quantitative description of the function of six genes. The success in applying this quantitative description to better understanding gene expression invites consideration of what other biological processes might benefit from gene-specific quantitative profiles. Improved understanding of the signaling properties of the BMP pathway have recently been possible through functional experiments titrating ratios of ligand pairs in the context of different heterodimeric receptor combinations (Antebi et al., 2017). The results of this study demonstrate that individual receptor pairs perform different “computations” as a function of the concentration of ligand pairs in the outside environment, with emergent preferences for either equal or unequal ligand ratios depending on the particular receptor subunits. Presumably, a deep understanding of how signaling pathways operate to shape the development of organisms will be advanced by quantitative frameworks such as these. Such a paradigm might also be applied to understanding metabolic pathways, to the extent that the nutrient sensing proteins and metabolic enzymes each recognize a variety of small-molecule substrates, with a range of binding affinities and catalytic rate constants. This will require high-throughput assays capable of producing quantitative profiles relevant to each gene product as well as mathematical models that sufficiently describe the known behavior of the relevant pathway. Indeed, such quantitative efforts will presumably become more apparently crucial to biological science as the wellspring of “undiscovered” gene functions becomes exhausted, wherein the remaining challenge will be to understand the complex, quantitative nature of interaction of these gene functions within cells, tissues, and entire organisms.

References

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, e05005.

Antebi, Y.E., Linton, J.M., Klumpe, H., Bintu, B., Gong, M., Su, C., McCardell, R., and Elowitz, M.B. (2017). Combinatorial signal perception in the BMP pathway. Cell *170*, 1184–1196.e24.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol. *18*, 285–298.

Braun, J.E., Huntzinger, E., Fauser, M., and Izaurralde, E. (2011). GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. Mol. Cell *44*, 120–133.

Braun, J.E., Huntzinger, E., and Izaurralde, E. (2012). A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. Cold Spring Harb. Perspect. Biol. *4*, a012328.

Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., and Gilbert, W.V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature *515*, 143–146.

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Nostrand, E.L.V., Pratt, G.A., et al. (2018). Sequence, structure, and context preferences of human RNA binding proteins. Mol. Cell *70*, 854–867.e9.

Eisen, T.J., Eichhorn, S.W., Subtelny, A.O., and Bartel, D.P. (2020). MicroRNAs cause accelerated decay of short-tailed target mRNAs. Mol. Cell *77*, 775–785.e8.

Elkayam, E., Faehnle, C.R., Morales, M., Sun, J., Li, H., and Joshua-Tor, L. (2017). Multivalent recruitment of human Argonaute by GW182. Mol. Cell *67*, 646–658.e3.

Eulalio, A., Huntzinger, E., and Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. Nat. Struct. Mol. Biol. *15*, 346–353.

Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2013). Rapid and specific purification of Argonaute-small RNA complexes from crude cell lysates. RNA *19*, 271–279.

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

Gosline, S.J.C., Gurtan, A.M., JnBaptiste, C.K., Bosson, A., Milani, P., Dalin, S., Matthews, B.J., Yap, Y.S., Sharp, P.A., and Fraenkel, E. (2016). Elucidating microRNA regulatory networks using transcriptional, post-transcriptional, and histone modification measurements. Cell Rep. *14*, 310–319.

Han, Y., Branon, T.C., Martell, J.D., Boassa, D., Shechner, D., Ellisman, M.H., and Ting, A. (2019). Directed evolution of split APEX2 peroxidase. ACS Chem. Biol. *14*, 619–635.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A phase separation model for transcriptional control. Cell *169*, 13–23.

Huberdeau, M.Q., Zeitler, D.M., Hauptmann, J., Bruckmann, A., Fressigné, L., Danner, J., Piquet, S., Strieder, N., Engelmann, J.C., Jannot, G., et al. (2017). Phosphorylation of Argonaute proteins affects mRNA binding and is essential for microRNA-guided gene silencing in vivo. EMBO J. *36*, 2088–2106.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

Larsson, E., Sander, C., and Marks, D. (2010). mRNA turnover rate limits siRNA and microRNA efficacy. Mol. Syst. Biol. *6*, 433.

Liu, J., Li, K., Cai, J., Zhang, M., Zhang, X., Xiong, X., Meng, H., Xu, X., Huang, Z., Peng, J., et al. (2019). Landscape and regulation of m6A and m6Am methylome across human and mouse tissues. Mol. Cell *77*, 426–440.e6.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. USA *108*, 11063–11068.

Luo, Y., Na, Z., and Slavoff, S.A. (2018). P-bodies: composition, properties, and functions. Biochemistry *57*, 2424–2431.

Macklin, D.N., Ahn-Horst, T.A., Choi, H., Ruggero, N.A., Carrera, J., Mason, J.C., Sun, G., Agmon, E., DeFelice, M.M., Maayan, I., et al. (2020). Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. Science *369*, eaav3751.

Nostrand, E.L.V., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719.

Padrón, A., Iwasaki, S., and Ingolia, N.T. (2019). Proximity RNA labeling by APEX-seq reveals the organization of translation initiation complexes and repressive RNA granules. Mol. Cell *75*, 875–887.e5.

Parker, R., and Sheth, U. (2007). P bodies and the control of mRNA translation and degradation. Mol. Cell *25*, 635–646.

Patel, R.K., West, J.D., Jiang, Y., Fogarty, E.A., and Grimson, A. (2020). Robust partitioning of microRNA targets from downstream regulatory changes. Nucleic Acids Res. *48*, gkaa687-.

Pfaff, J., Hennig, J., Herzog, F., Aebersold, R., Sattler, M., Niessing, D., and Meister, G. (2013). Structural features of Argonaute–GW182 protein interactions. Proc. Natl. Acad. Sci. USA *110*, E3770–E3779.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., et al. (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111–1124.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Sharma, D., Zagore, L.L., Brister, M.M., Ye, X., Crespo-Hernández, C.E., Licatalosi, D.D., and Jankowsky, E. (2020). The kinetic landscape of an RNA binding protein in cells. bioRxiv 2020.05.11.089102.

Zaccara, S., Ries, R.J., and Jaffrey, S.R. (2019). Reading, writing and erasing mRNA methylation. Nat. Rev. Mol. Cell Biol. *20*, 608–624.

Zubradt, M., Gupta, P., Persad, S., Lambowitz, A.M., Weissman, J.S., and Rouskin, S. (2017). DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat. Methods *14*, 75–82.

Appendix A.

Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing

Vincent C. Auyeung[1,2,3,4], Igor Ulitsky[1,2,3], Sean E. McGeary[1,2,3], and David P. Bartel[1,2,3,4]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02139, USA

Published as:
Auyeung, V.A., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell *152*, 844–858.

Cell

Beyond Secondary Structure: Primary-Sequence Determinants License Pri-miRNA Hairpins for Processing

Vincent C. Auyeung,[1,2,3,4] Igor Ulitsky,[1,2,3] Sean E. McGeary,[1,2,3] and David P. Bartel[1,2,3,*]
[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA 02139, USA
*Correspondence: dbartel@wi.mit.edu
http://dx.doi.org/10.1016/j.cell.2013.01.031

SUMMARY

To use microRNAs to downregulate mRNA targets, cells must first process these ~22 nt RNAs from primary transcripts (pri-miRNAs). These transcripts form RNA hairpins important for processing, but additional determinants must distinguish pri-miRNAs from the many other hairpin-containing transcripts expressed in each cell. Illustrating the complexity of this recognition, we show that most *Caenorhabditis elegans* pri-miRNAs lack determinants required for processing in human cells. To find these determinants, we generated many variants of four human pri-miRNAs, sequenced millions that retained function, and compared them with the starting variants. Our results confirmed the importance of pairing in the stem and revealed three primary-sequence determinants, including an SRp20-binding motif (CNNC) found downstream of most pri-miRNA hairpins in bilaterian animals, but not in nematodes. Adding this and other determinants to *C. elegans* pri-miRNAs imparted efficient processing in human cells, thereby confirming the importance of primary-sequence determinants for distinguishing pri-miRNAs from other hairpin-containing transcripts.

INTRODUCTION

MicroRNAs (miRNAs) are ~22 nt RNAs that pair to messenger RNAs (mRNAs) to direct posttranscriptional repression (Bartel, 2004). MicroRNAs are processed from hairpin-containing primary transcripts (pri-miRNAs). In the canonical processing pathway of animals, pri-miRNAs are cleaved by the Microprocessor, a protein complex containing an RNase III enzyme, Drosha, and its cofactor, DGCR8/Pasha (Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). The liberated portion of the hairpin (the pre-miRNA) is then cleaved by the RNase III enzyme Dicer (Grishok et al., 2001; Hutvágner et al., 2001), leaving two ~22 nt strands that pair to each other with ~2 nt 3′ overhangs (Lee et al., 2003; Lim et al., 2003b). One strand of each duplex is loaded into an Argonaute protein to form the core of the silencing complex, and the other strand is discarded (Khvorova et al., 2003; Schwarz et al., 2003; Liu et al., 2004). Noncanonical pathways also contribute to the miRNA repertoire through the processing of mirtrons (Okamura et al., 2007; Ruby et al., 2007) or other pri-miRNAs that bypass Drosha cleavage (Babiarz et al., 2008) and through one pre-miRNA that bypasses Dicer cleavage (Cheloufi et al., 2010; Cifuentes et al., 2010).

A long-standing mystery has been how pri-miRNAs are distinguished from the many other hairpin-containing transcripts for processing as Microprocessor substrates. Determinants of Dicer cleavage are better understood (Zhang et al., 2004; Macrae et al., 2006; Park et al., 2011), as illustrated by both the design (Brummelkamp et al., 2002; Paddison et al., 2002) and prediction (Chung et al., 2011) of Dicer substrates that bypass Drosha processing. For Microprocessor recognition, sequences within 40 nt upstream and 40 nt downstream of the pre-miRNA hairpins are required for ectopic miRNA expression (Chen et al., 2004), which is consistent with (1) the observation that these flanking sequences tend to pair to each other to extend the stem another turn of the helix beyond the cleavage site (Lim et al., 2003b) and (2) a requirement for both this extension and a lack of pairing immediately following it for processing (Han et al., 2006). However, many cellular transcripts have paired regions flanked by single-stranded RNA (ssRNA), and most of these are not Microprocessor substrates. Indeed, attempts to predict canonical miRNA hairpins from genomic sequence yield many thousands of false-positive predictions, which must be eliminated using additional criteria, such as analysis of conservation or experimental evaluation (Lim et al., 2003a, 2003b; Bentwich et al., 2005; Berezikov et al., 2006; Chiang et al., 2010). This illustrates a large gap in our understanding of how the Microprocessor distinguishes between authentic substrates and other transcribed hairpins.

Here, we report that transcripts that enter the miRNA pathway in *C. elegans* failed to do so in human cells. Thus, the definition of a pri-miRNA in one species differs from that in another.

To find features that define human pri-miRNAs, we generated more than 10^{11} variants of four pri-miRNAs and sequenced millions that were cleaved by the human Microprocessor. Comparison of cleaved and initial variants revealed important sequence and structural features. These features were evolutionarily conserved in non-nematode lineages and sufficient to increase the processing efficiency of *C. elegans* hairpins in human cells.

RESULTS

Unknown Features Specify Human Pri-miRNAs

To examine whether miRNA processing features are shared across animals, we ectopically expressed a panel of *C. elegans*, *D. melanogaster*, and human pri-miRNAs in human cells and compared the yields of mature miRNA. Despite variability in the degree of overexpression, presumably reflecting differences in efficiency at various steps of the pathway (Fellmann et al., 2011; Feng et al., 2011), most human miRNAs were efficiently expressed (Figure 1A), as expected (Chiang et al., 2010). Four of nine *Drosophila* miRNAs also fell within the range observed for human miRNAs. However, the tested *C. elegans* miRNAs were less efficiently expressed (Figure 1A, $p = 1.4 \times 10^{-5}$, Wilcoxon rank-sum test). Similar results were observed in *Drosophila* S2 cells ($p = 0.024$). Thus, most nematode pri-miRNAs lack determinants required for efficient processing in human or insect cells.

To isolate the processing defect, we probed for processing intermediates. Consistent with the sequencing results, *cel-lin-4* was processed, with detectable pre-miRNA and mature miRNA (Figure 1B). For other *C. elegans* miRNAs, neither pre-miRNA nor mature miRNA was detected, despite the presence of primary transcripts (Figure 1B; Figure S1B, available online), suggesting that these *C. elegans* pri-miRNAs were not productively recognized as Microprocessor substrates. To assay directly for Microprocessor binding, we examined binding to catalytically deficient Drosha and DGCR8. Whereas human *pri-mir-122* bound the Microprocessor somewhat better than did the reference pri-miRNA (human *pri-mir-125a*), all seven tested *C. elegans* pri-miRNAs bound worse (Figure 1C). Thus, most *C. elegans* pri-miRNAs are missing some of the determinants needed for efficient recognition and processing by the human Microprocessor.

Known features of *C. elegans* and human pri-miRNAs appear largely similar, as illustrated by the accuracy of an algorithm trained on *C. elegans* pri-miRNAs in predicting most miRNA genes conserved in mammals and fish (Lim et al., 2003a). Nonetheless, the poor specificity of this algorithm when predicting nonconserved miRNAs suggests that unknown features help define authentic pri-miRNAs. To look for clues regarding these unknown features, we analyzed the conservation of sequence immediately flanking human pre-miRNAs. Residues extending 13 nt upstream of the 5p Drosha cleavage site (i.e., the site corresponding to the 5′ end of the pre-miRNA) and 11 nt downstream of the 3p Drosha cleavage site were conserved above background, consistent with the importance of the ~11 bp basal stem for pri-miRNA processing (Figure 1D). However, the signal beyond the basal stem tailed off rapidly (particularly in the upstream flanking region), suggesting that any determinants in the flanking regions might be either at variable distances from the hairpin or present in only subsets of miRNAs, making them difficult to identify using alignments.

Functional Substrates from Large Libraries of Pri-miRNA Variants

To identify features important for Microprocessor recognition and cleavage, we generated more than 10^{11} pri-miRNA variants, sequenced millions that retained function, and compared these sequences to those of the initial variants (Figure 2A). This approach resembled classical in vitro selection approaches (Wilson and Szostak, 1999), except we did not perform multiple rounds of selection. Because the starting and the selected pools underwent the same number of transcription, reverse-transcription, and amplification steps, any differences between the two pools were subject to neither the compounding effects of multiple rounds nor the confounding effects of amplification biases. Moreover, as with previous analyses of selection results using high-throughput sequencing (Zykovich et al., 2009; Pitt and Ferré-D'Amaré, 2010; Slattery et al., 2011), sequencing depth reduced the influence of stochastic sampling. Thus, compared to the results of classical approaches, enrichment or depletion of a residue was a more direct reflection of its contribution to biochemical specificity.

Four pools of variants were constructed, each based on a different human pri-miRNA (*mir-125a*, *mir-16-1*, *mir-30a*, and *mir-223*). Residues more than 8 nt upstream of the 5p Drosha cleavage site and more than 8 nt downstream of the 3p cleavage site were varied, whereas the remaining hairpin residues were not. At each variable position, 79% of the molecules had the wild-type residue, and the remainder had one of the other three alternatives. As done for self-cleaving ribozymes (Pan and Uhlenbeck, 1992), each variant was circularized so that all of its variable nucleotides resided in a single cleavage product (Figure 2A), thereby enabling a full analysis of sequence interdependencies.

In vitro cleavage reactions were conducted in Microprocessor lysate, i.e., whole-cell lysate from HEK293T cells overexpressing Drosha and DGCR8 to enhance cleavage activity (Figure 2B). At a time in which the lysate cleaved linear and circularized *pri-mir-125a* to near completion, many *pri-mir-125a* variants remained uncleaved (Figure 2C), which indicated that some substitutions in the basal stem and flanking regions attenuated Microprocessor cleavage in vitro.

Cleaved variants were purified and sequenced (Figure 2A). At each variant position, the odds of each nucleotide in the cleaved pool were compared to the odds of that nucleotide in the starting pool. These odds ratios were used to calculate the information content of each nucleotide possibility at each variant position—the greater the information content, the more favorable the influence on activity, with positive values indicating beneficial influences and negative values disruptive ones. An advantage of plotting information content is that it reports the relative influence of each nucleotide possibility irrespective of whether it was the wild-type possibility. Because molecular manipulations and computational filtering both selected for cleavage at the wild-type site, nucleotide changes

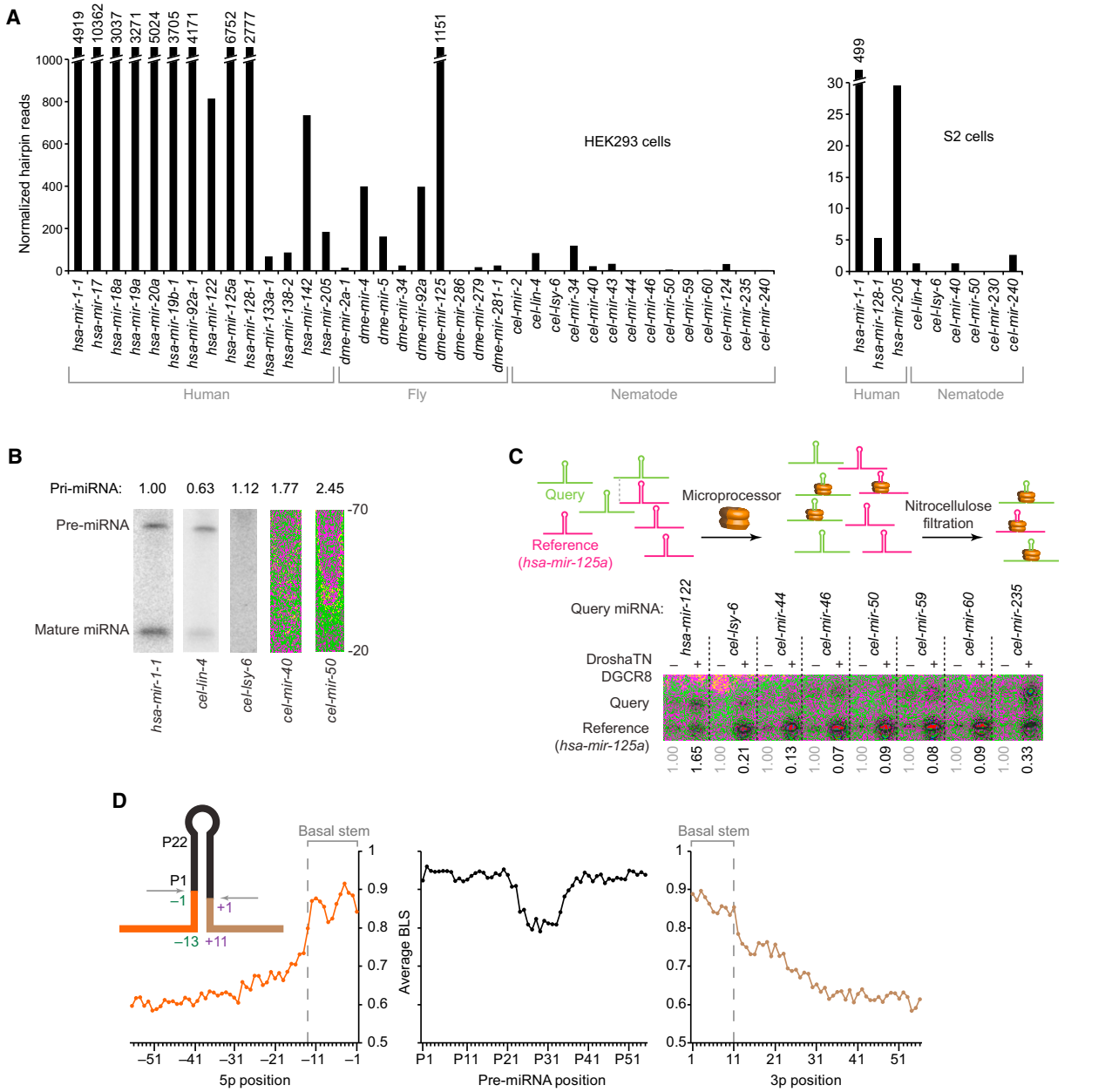


Figure 1. Existence of Unknown Features Specifying Human Pri-miRNAs

(A) Processing of human, fly, and nematode pri-miRNAs in human cells and *Drosophila* cells. Cells were transfected with plasmids expressing the indicated pri-miRNA hairpins with ~100 flanking genomic nucleotides on each side of each hairpin (Figure S1A), and total RNA was pooled for small-RNA sequencing. Plotted are small-RNA reads derived from the indicated pri-miRNAs.

(B) Accumulation of pri-miRNA, pre-miRNA, and miRNA after expressing the indicated pri-miRNAs in HEK293T cells. Pre-miRNA and mature species were measured by RNA blot of total RNA from cells transfected with plasmids expressing the indicated pri-miRNA (full gel images, including in vitro-transcribed cognate positive controls, in Figure S1B). Relative pri-miRNA levels (indicated above the lanes) are from ribonuclease protection assays, normalized to the signals for neomycin phosphotransferase mRNA also expressed from each expression plasmid.

(C) Relative binding of *C. elegans* and human pri-miRNAs to the Microprocessor. In the competitive binding assay (top, schematic), radiolabeled query pri-miRNA was mixed with the radiolabeled shorter reference pri-miRNA (human *mir-125a*) and incubated in excess over catalytically impaired Drosha (Drosha-TN) and DGCR8. Bound RNA was filtered on nitrocellulose and eluted for analysis on a denaturing gel. Phosphorimaging (bottom) indicated the relative amounts of input (–) and bound (+) RNAs. Numbers below each lane indicate the ratio of bound query to bound reference pri-miRNAs, normalized to their input ratio.

(D) Nucleotide conservation of human pri-miRNAs conserved to mouse, reported as the average branch-length score (BLS) at each position. Positions are numbered based on the inferred Drosha cleavage site (inset); negative indices are upstream of the 5p Drosha cleavage site, indices with "P" count from the 5′ end of the pre-miRNA, and positive indices are downstream of the 3p Drosha cleavage site.

See also Figure S1.

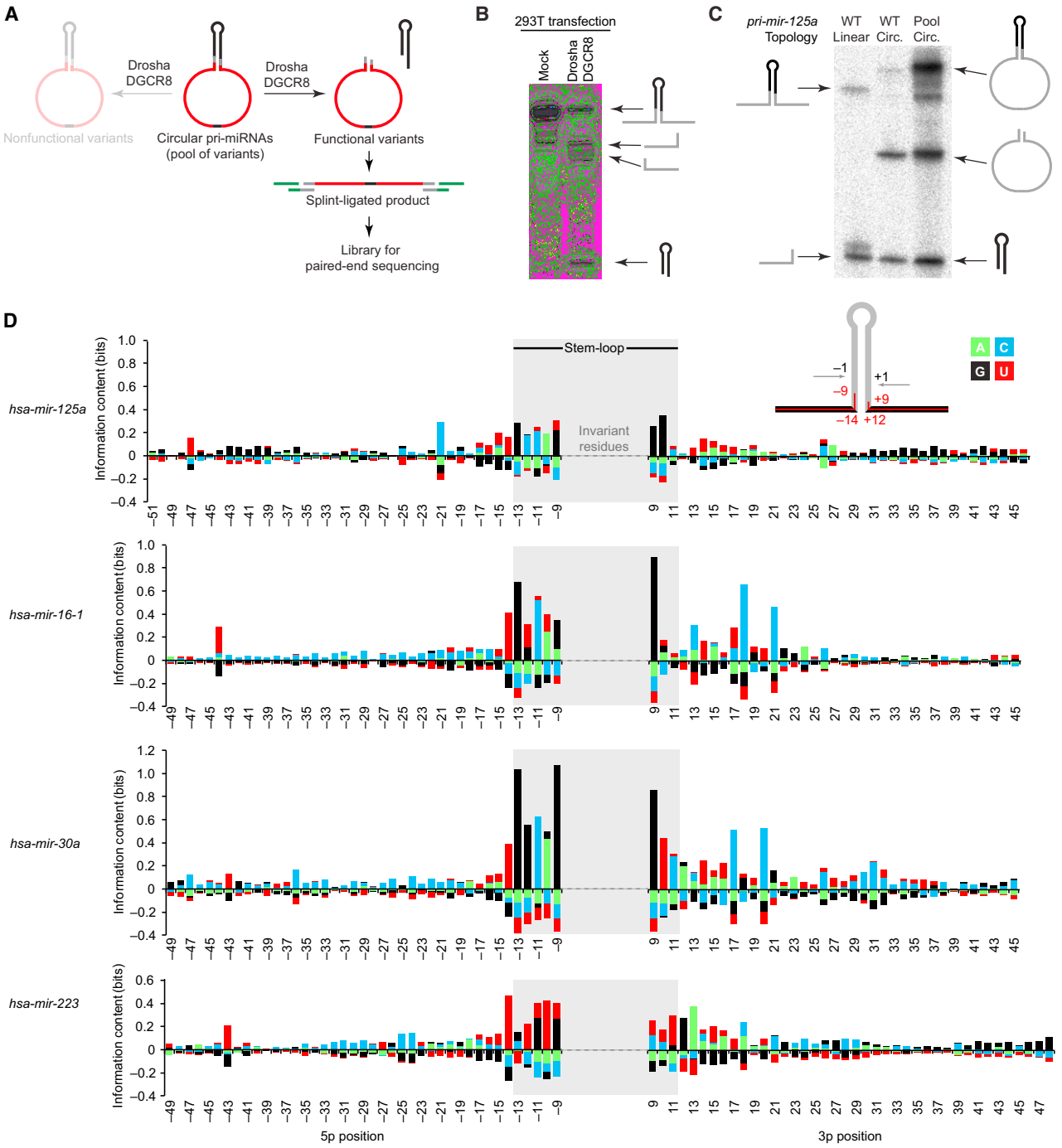


Figure 2. Selection for Functional Pri-miRNA Variants

(A) Schematic of the selection. Pri-miRNAs with variable residues (red) flanking the Drosha cleavage site were circularized by ligation and incubated in Microprocessor lysate. Cleaved variants were gel purified, ligated to adaptors, reverse transcribed, and amplified for high-throughput sequencing.

(B) Cleavage of *let-7a* in HEK293T whole-cell lysate (mock) and Microprocessor lysate (whole-cell lysate from HEK293T cells transfected with plasmids expressing Drosha and DGCR8). Incubations were 1.5 hr. Body-labeled reactants and products were resolved on a denaturing polyacrylamide gel and visualized by phosphorimaging.

(C) Cleavage of linear and circular *mir-125a* (WT linear and WT circ., respectively) and a pool of circular *mir-125a* variants (pool). RNAs were incubated for 5 min in Microprocessor lysate and analyzed as in (B). The linear RNA was 5′ end labeled; other RNAs were body labeled.

(D) Enrichment and depletion at variable residues in functional pri-miRNA variants. At each varied position (inset, red inner line), information content was calculated for each residue (green, cyan, black, and red for A, C, G, and U, respectively).

See also Figure S2.

that altered the cleavage site were not distinguished from those that abolished cleavage.

Some positions had substantial enrichment of one or more nucleotide possibilities, with corresponding depletion of others (Figure 2D). When tested in vitro, the results of changing specific residues closely matched those predicted from analysis of sequenced variants (Figures S2A and S2B). Moreover, the in vitro results predicted the direction and sometimes the magnitude of the effects observed in HEK293T cells (Figure S2C).

Importance of an 11 bp Basal Stem Flanked by at Least Nine Unstructured Nucleotides

For all four miRNAs, some of the varied residues with the greatest influence fell within the basal stem (Figure 2D). Covariation matrices listing the odds ratio of each pair of nucleotide identities showed preference for Watson-Crick geometry at each basal pair, with the G:U wobble the most frequently preferred non-Watson-Crick alternative (Figures 3A and S3A). For example, the most favored alternatives to the wild-type C:G pair at positions −11 and +9 of *mir-125a* were the G:C and U:A pairs, and to a lesser extent, the A:U, G:U, and U:G pairs (Figure 3A). In fact, Watson-Crick pairing was strongly preferred even if it did not occur in the wild-type sequence. For example, the wild-type A:C pair at positions −12 and +10 of *mir-30a* was disfavored compared to the four Watson-Crick possibilities (Figure 3A), and the bulged A at position +10 of *mir-223* was preferentially incorporated into an alternative continuous helix (Figures S3A and S3B). Extending these methods to systematically evaluate all pairing possibilities involving all varied positions uncovered no evidence for Watson-Crick pairing outside the basal stem (Figure S3C).

Layered on the overall preference for Watson-Crick pairing were primary-sequence preferences specific to each basal pair. For example, at positions −11 and +9 the C:G pair was favored over the other Watson-Crick alternatives. The primary-sequence preference was most acute at the basal-most pair, where wobbles or mismatches involving G at −13 were favored over alternative Watson-Crick pairs (Figure 3A). We conclude that primary-sequence features supplement and sometimes supersede structural features important for basal-stem recognition.

The Microprocessor recognizes the junction between the miRNA hairpin and flanking ssRNA to position the active site approximately one helical turn (11 bp of A-form RNA) from the base of the duplex (Han et al., 2006; Yeom et al., 2006). To examine the preferred length of the basal stem, we calculated the relative cleavage efficiencies of different stem-length variants, normalizing to that of an 8 bp stem. Invariant mismatches within symmetric internal loops (e.g., the A:C mismatch at positions −6 and +4 of *mir-30a*) were assumed to be noncanonical pairs that stacked within the stem to contribute to its length, whereas mismatches at varied positions were assumed to disrupt further pairing and thereby terminate the inferred basal stem. For all four pri-miRNAs, an 11 bp basal stem was optimal (Figure 3B), consistent with the single-turn model. Indeed, an 11 bp basal stem was preferred for *mir-223* even though the wild-type sequence was predicted to form a 12 bp stem (Figures 3A and S3A). For most pri-miRNAs, however, the efficiency of the 12-pair stem approached that of the 11-pair stem (Figure 3B). This tolerance of a twelfth pair hinted that other features, such as the G at position −13, help specify the precise site of cleavage.

The single-turn model also posits that the nucleotides immediately flanking the basal stem are unstructured (Han et al., 2006; Yeom et al., 2006). To test this, we used RNAfold (Hofacker and Stadler, 2006) to predict the minimum free-energy structure of each sequenced pri-miRNA variant. For those with predicted wild-type stem pairing, we recorded the number of nucleotides between the base of the stem and the most proximal two consecutive structured residues. Although an imperfect estimate of the size of the unstructured segments flanking the base of the helix, this metric correlated well with cleavage (Figure 3C). Predicted pairing was tolerated in one flank, provided that the other flank contained at least 5–7 unpaired bases, consistent with reports of some cleavage when only one flanking segment is present (Zeng and Cullen, 2005; Han et al., 2006). When summing the flanking unpaired bases from both sides, the optimum plateaued at ~9–18 nt (Figure 3D).

A Basal UG Motif Enhances Processing

Among the nucleotides upstream of the stemloop, the most striking enrichment was for a U at position −14 (Figure 2D). This U immediately preceded the position that, as mentioned above, displayed a strong primary-sequence preference for a G. The U and G at positions −14 and −13 contributed independently; variants with either a U or a G were enriched over variants with neither, and variants with both were even more enriched (Figure 4A). For *mir-223*, the UG at positions −14 and −13 was preferred (Figure 2D), even though wild-type *mir-223* has a UG at positions −15 and −14, respectively. This basal UG motif was also enriched among variants of *mir-125a* selected for Microprocessor binding rather than cleavage (Figure S4B).

The basal UG was conserved in vertebrate orthologs of *mir-16-1* and *mir-30a* (Figure 4B). Moreover, the motif was enriched in other mammalian pri-miRNAs, as illustrated by the sequence composition of human pri-miRNAs (Figure 4C). It was also enriched in pri-miRNAs of zebrafish (*D. rerio*) and tunicate (*C. intestinalis*) but only sporadically in more distantly related lineages, suggesting that its recognition emerged in a chordate ancestor (Figure 4D).

The Broadly Conserved CNNC Motif Enhances Processing

In *mir-16-1*, *mir-30a*, and *mir-223* we observed a preference for two C residues, separated by two intervening nucleotides, beginning 17–18 nt downstream of the Drosha cleavage site (Figure 2D). The two C residues of this CNNC motif (N signifies any nucleotide) acted synergistically, in that variants that retained neither C residue were not disfavored much more than those that retained one (Figure 5A). The C residues enriched in the active variants were conserved in vertebrate orthologs of these three pri-miRNAs (Figure 5B).

The *mir-125a* pri-miRNA also had four C residues in this vicinity (positions 16–21), which gave rise to a CNNC at position

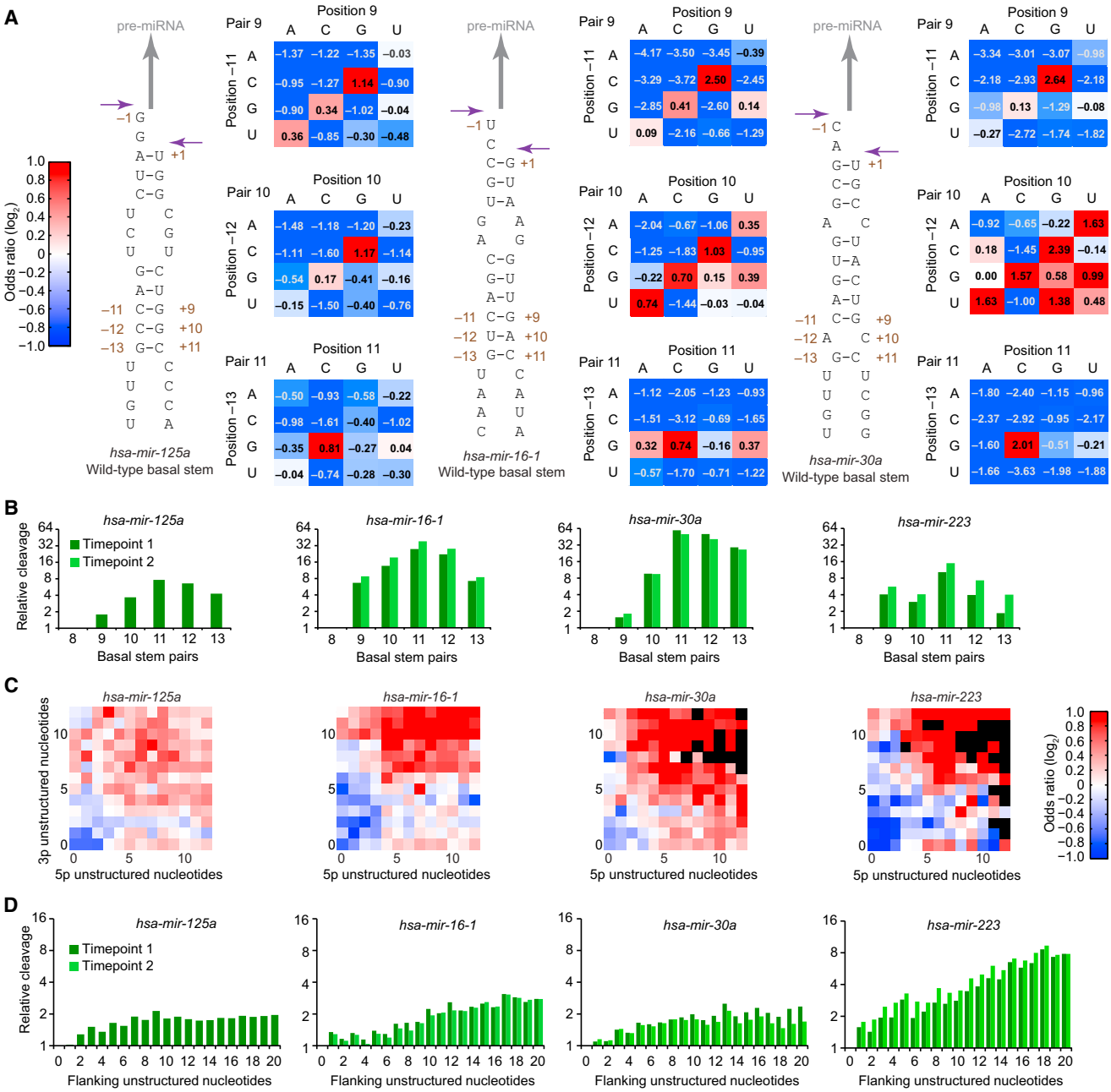


Figure 3. Basal Stem Structure in Functional Pri-miRNA Variants

(A) Predicted basal secondary structures and covariation matrices for *mir-125a*, *mir-16-1*, and *mir-30a*. For each pair of positions, joint nucleotide distributions were tabulated from sequences of the initial and selected pools, and the log odds ratio was calculated. Favored and disfavored pairs are colored red and blue, respectively, with color intensity (key) and values indicating magnitudes.

(B) Relative cleavage of variants with different stem lengths. The number of contiguous Watson-Crick pairs was counted, and the relative cleavage was calculated, normalized to the 8 bp stem. For selections with two time points, results are shown for both (key).

(C) Enrichment for unstructured nucleotides flanking the basal stem. Predicted folds of variant sequences were generated, and the subset of sequences with wild-type basal stem pairing were classified based on the distance to the nearest consecutive structured nucleotides upstream of position −13 and the nearest consecutive structured nucleotides downstream of position +11. Enrichment (red) and depletion (blue) of unstructured lengths among the selected variants are colored (key), with black indicating that sequencing data were insufficient to calculate enrichment.

(D) Relative cleavage of variants with differing numbers of total unstructured nucleotides flanking the basal stem. Upstream and downstream unstructured lengths predicted in (C) were summed, and the relative cleavage was calculated, normalized to zero unstructured nucleotides. For selections with two time points, results are shown for both (key).

See also Figure S3.

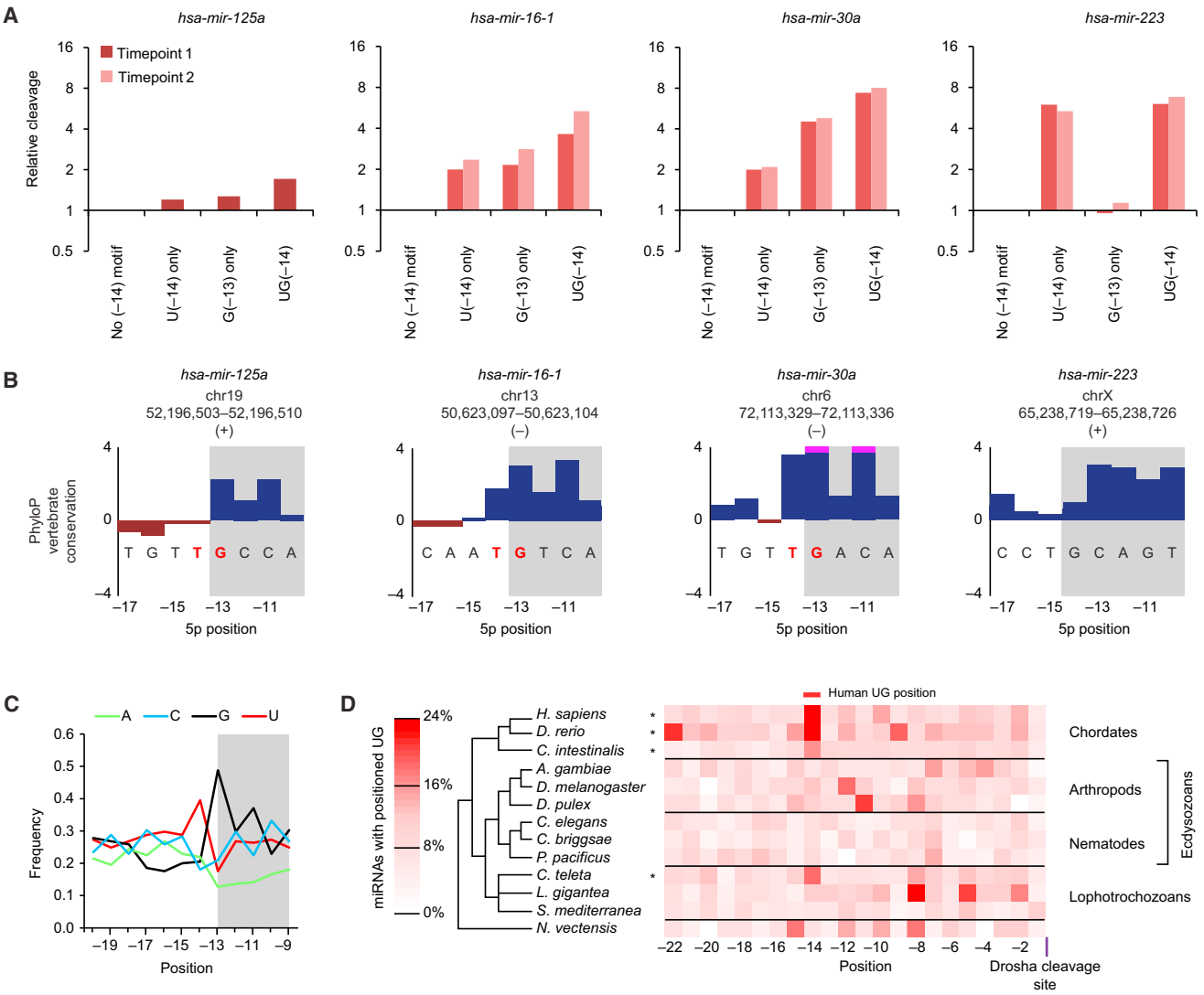


Figure 4. The Basal UG Motif

(A) Relative cleavage of variants with a full UG motif, a partial motif, and no motif. Values were normalized to those of variants with no motif, showing results from two time points, if available (key).

(B) PhyloP conservation across 30 vertebrate species in the region of the basal UG motif (red letters) for the four selected miRNAs. Bars extending beyond the scale of the graph are truncated (pink). Nucleotides predicted to be paired in the wild-type basal stem are shaded.

(C) Frequencies of A, C, G, and U (green, cyan, black, and red, respectively) at the indicated positions of human pri-miRNAs conserved to mouse. Analysis was of 204 pri-miRNAs, each representing a unique paralogous family (Table S2).

(D) Enrichment for the UG dinucleotide in the pri-miRNAs of representative animals with sequenced genomes. UG occurrences were tabulated for the upstream regions of pri-miRNAs aligned on the predicted Drosha cleavage site (Table S2). Species with statistically significant enrichment at position −14 are indicated (asterisks, empirical p value < 10^{-3}).

See also Figure S4.

16 and the possibility of creating a CNNC at positions 17 or 18 (by changing either A20 or A18, respectively, to a C). However, the CNNC at position 16 was not preferred in the selection, nor were either of the single-nucleotide changes that could create a CNNC (Figures 2D and 5A). Moreover, the position 16 CNNC was not conserved in vertebrate orthologs (Figure 5B). These results indicate that unidentified features present in *mir-16-1*, *mir-30a*, and *mir-223*, but not *mir-125a*, are required for the CNNC to increase processing efficiency.

For the three pri-miRNAs in which the CNNC motif was effective, its position fell in a small window 17–18 nt downstream of the Drosha cleavage site. In variants in which neither wild-type C was present, alternative CNNC motifs were strongly enriched 1–2 nt downstream (Figure S5A), which further indicated that a CNNC motif within a small range of positions can contribute to pri-miRNA recognition.

Of the 64 possible dinucleotide motifs with zero to three intervening nucleotides, CNNC was the one most highly enriched

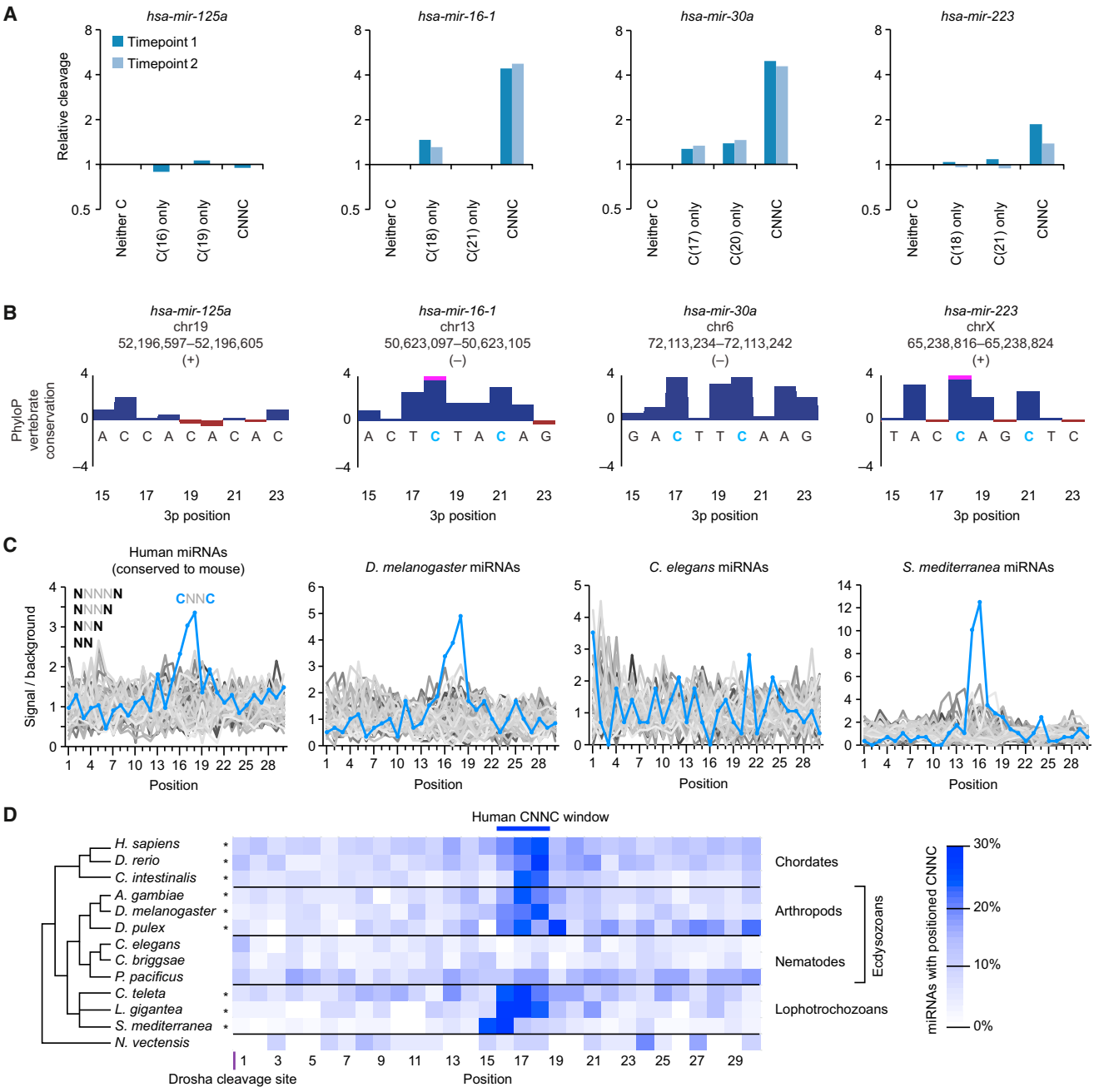


Figure 5. The Downstream CNNC Motif

(A) Relative cleavage of variants with a full CNNC motif, a partial motif, and no motif. Values were normalized to those of variants with no motif, showing results from two time points, if available (key).

(B) PhyloP conservation across 30 vertebrate species in the region of the downstream CNNC motif (blue letters) for the four selected pri-miRNAs. Bars extending beyond the scale of the graph are truncated (pink).

(C) CNNC enrichment compared to that of 63 other spaced dinucleotide motifs. Occurrences of each motif were tabulated for the downstream regions of pri-miRNAs aligned on the predicted Drosha cleavage site (Table S2). Background expectation was based on the nucleotide composition of pri-miRNA downstream regions in each species.

(D) Enrichment of the CNNC motif in the pri-miRNAs of representative bilaterian animals (Table S2). Species with statistically significant enrichment at positions 16, 17, or 18 are indicated (asterisk, empirical p value < 10^{-4}).

See also Figure S5.

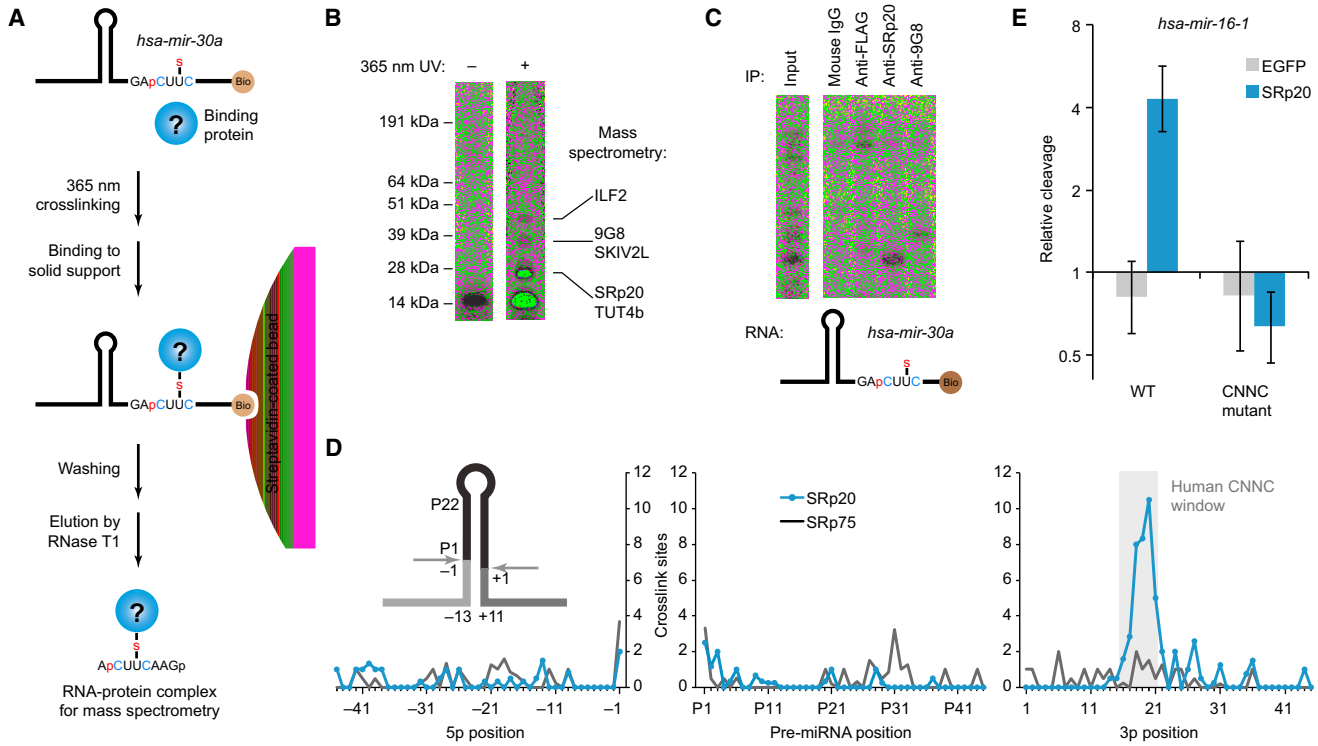


Figure 6. Binding and Activity of SRp20 at the CNNC Motif

(A) Site-specific crosslinking approach used to identify CNNC-binding proteins. The *mir-30a* crosslinking substrate contained a photoreactive base in the CNNC motif (4-thiouridine, U–S), a 3′ biotin (Bio), and for some applications, a ^{32}P-labeled phosphate (red p). This substrate was incubated in Microprocessor lysate and irradiated with 365 nm UV light. Crosslinked complexes were captured on streptavidin-coated beads and eluted by RNase T1 digestion.

(B) Proteins within crosslinked RNA-protein complexes. Crosslinked complexes prepared as in (A) were separated on an SDS gel. For each CNNC-crosslinked band, proteins are listed that were identified by mass spectrometry and have known or inferred RNA-binding activity.

(C) Immunoprecipitation of proteins crosslinked to the CNNC motif. After crosslinking as in (A), complexes were enriched using monoclonal antibodies against either FLAG (the tag of the overexpressed Drosha and DGCR8), SRp20, or 9G8 and then resolved on an SDS gel. Input was run on a different region of the same gel for reference.

(D) SRp20 binding downstream of mouse pri-miRNA hairpins in vivo. Sites were obtained by reanalysis of crosslinking data for SRp20 and SRp75 in mouse cells (Änkö et al., 2012). Positions are numbered as in Figure 1D. Expected sites of crosslinks to any of the motif nucleotides in the region of motif enrichment (Figure 5D) are shaded (gray).

(E) Enhancement of in vitro pri-miRNA cleavage by SRp20. Wild-type *pri-mir-16-1* or *pri-mir-16-1* with mutated CNNC were incubated for 3 min with immunopurified Microprocessor, supplemented with either FLAG-EGFP or 3X-FLAG-SRp20 purified from HEK293T cells. Reactants and products were resolved on denaturing polyacrylamide gels and quantified by phosphorimaging relative to a buffer-only control (geometric mean ± standard error, n = 3).

See also Figure S6.

downstream of the cleavage sites of human pri-miRNAs (Figure 5C). Moreover, enrichment was limited to a small range of positions 16–18 nt downstream of the site, peaking at positions 17 and 18, which matched the positions of the motif within *mir-16-1*, *mir-30a*, and *mir-223*. These results suggest that the CNNC motif enhances processing of many human pri-miRNAs.

Similar analyses of nonmammalian pri-miRNAs indicated strong, position-specific enrichment of the CNNC motif in chordates, arthropods, and lophotrochozoans, but not in sea anemone (*Nematostella vectensis*) (Figures 5C and 5D), suggesting that its recognition emerged with the divergence of bilaterians. Interestingly, enrichment was also absent in nematodes (Figures 5C and 5D), suggesting an isolated loss in the nematode branch of the ecdysozoans.

Consistent with the results in extracts, mutation of the basal UG and downstream CNNC motifs each reduced accumulation of mature miR-16 and miR-30a in HEK293T cells, with mutation of both reducing accumulation ~4–8-fold relative to wild-type (Figures S5B and S5C). Furthermore, one or both motifs contributed to the accumulation of each of the additional pri-miRNAs tested in cell culture (*hsa-mir-28*, *hsa-mir-129-2*, and *hsa-mir-193b*; Figures S5D–S5F).

SRp20 Binds the CNNC Motif and Enhances Processing

To learn how the CNNC motif is recognized, we used site-specific crosslinking (Wyatt et al., 1992). Proteins that crosslinked to *pri-mir-30a* RNA with a photoreactive nucleotide (4-thiouridine) placed within the CNNC motif were identified by mass spectrometry (Figure 6A). To guide gel-purification of

crosslinked proteins, we performed the procedure in parallel with a radiolabeled pri-miRNA designed to label only proteins that crosslinked in the vicinity of the CNNC (Figures 6A and 6B). The two strongest candidates were SRp20/SRSF3 and 9G8/SRSF7, closely related proteins implicated in splicing regulation (Zahler et al., 1993; Cavaloc et al., 1994), mRNA export (Huang and Steitz, 2001), and translation initiation (Bedard et al., 2007; Swartz et al., 2007). These proteins both have an RNA-recognition motif (RRM) conserved across bilaterian animals, which recognizes degenerate motifs closely related to the CNNC motif (Heinrichs and Baker, 1995; Cavaloc et al., 1999; Schaal and Maniatis, 1999). NMR studies of this RRM in complex with RNA indicate that the C residues, particularly the first C of the CNNC, are bound in a base-specific manner, with minimal preferences for the two intervening bases (Hargous et al., 2006). Immunopurification of SRp20 and 9G8 confirmed that these two proteins (particularly SRp20) were the ones that most efficiently crosslinked in our assay (Figure 6C).

To evaluate SRp20 binding in vivo, we analyzed a large data set of SRp20 crosslinking sites in P19 cells (Änkö et al., 2012). Although the published analyses of this data set focused on sites within pre-mRNAs, we found that many SRp20 sites resided in pri-miRNAs, and, more importantly, that these sites overlapped the region of CNNC enrichment (Figure 6D). This analysis extended our results from in vitro binding to in vivo binding and from one pri-miRNA to many. Some of the crosslinking sites in the CNNC-enriched region were in pri-miRNAs that lacked a CNNC motif, suggesting that SRp20 (and presumably its paralog, 9G8) might play a role even more general than that implied by CNNC conservation and enrichment.

The requirement of SRp20 for cell viability (Jumaa et al., 1999; Jia et al., 2010) confounded attempts to test its function by depleting the protein in cell culture. Therefore, we tested its function in vitro, supplementing immunopurified Microprocessor complex with either immunopurified recombinant SRp20 (Figure S6) or an analogously purified control protein (EGFP). SRp20 enhanced *mir-16-1* processing in a CNNC-dependent manner (Figure 6E). Taken together, our results indicate that for many bilaterian miRNAs the CNNC motif is enriched and preferentially conserved because it helps recruit SRp20 (or its homologs), which enhances pri-miRNA recognition and processing.

Loop and Apical Stem Elements Can Enhance Processing

To examine whether additional processing features reside in the loop and apical stem, we extended our approach to those regions (Figure S7A). Pairing at the apical portion of the stem contributed to pri-miRNA recognition and processing for *mir-125a* and *mir-30a*, but not for *mir-16-1* or *mir-223* (Figure S7B), consistent with differing conclusions drawn from studies of different miRNAs (Zeng et al., 2005; Han et al., 2006). Primary-sequence preferences were weaker than those observed for basal and flanking residues (Figure S7C). The best candidate for a loop-binding motif was observed only in *mir-30a*, in which the wild-type UGUG at positions P24–27 was both preferred in the selection (Figure S7D) and conserved in vertebrate orthologs (Figure S7E). Human and zebrafish miRNAs were enriched for UGU or GUG in this region of the loop (empirical $p < 10^{-5}$ for each species) (Figure S7F), thereby confirming it as the third primary-sequence motif identified in our study (Figure 7A).

Rescue of *C. elegans* miRNA Expression in Human Cells

The primary-sequence motifs important for mammalian miRNAs were not enriched in the nematode clade, suggesting that their absence might account for the failure of *C. elegans* pri-miRNAs to be processed in human cells. To test this idea, we added the basal UG and the downstream CNNC motifs to *cel-mir-44* in the context of the *mir-1* bicistronic vector (Figure 7B). Before adding the motifs, we disrupted the predicted pairing between positions −14 and +12 and substituted the G:C pair at positions −13 and +11 (construct mir44.1). These changes, which were expected to simultaneously enhance processing by shortening the basal stem to its optimal length and inhibit processing by replacing the fortuitous G at position −13, had a marginal net effect on production of mature miR-44 in human cells (Figure 7B). Adding a basal UG enhanced production of mature miR-44 by 5-fold (8-fold over the wild-type), primarily from restoring the G at −13 (Figure 7B). Adding a CNNC 17 nt downstream of the cleavage site (mir44.4) enhanced production another 8-fold, yielding a 64-fold net increase over wild-type (Figure 7B). Similarly, converting the wild-type, asymmetrically bulged stem of *cel-mir-50* to a regular, 11-pair stem and adding the UG and CNNC motifs enhanced expression of mature miR-50 by 30-fold (Figure S7G), while adding the motifs to *cel-mir-40* enhanced expression of mature miR-40 by 5-fold (Figure S7H). We conclude that primary-sequence motifs discovered in this study help human cells to distinguish pri-miRNA hairpins from other hairpins and that the absence of these motifs in *C. elegans* pri-miRNAs helps to explain why human cells do not regard these transcripts as pri-miRNAs.

DISCUSSION

Secondary structure is inadequate on its own to specify pri-miRNA hairpins: primary-sequence features, including the basal UG, the CNNC, and the apical GUG motifs, also contribute to efficient processing in human cells (Figure 7A). Complicating the story (and perhaps explaining why these primary-sequence features had not been observed earlier), different pri-miRNAs differentially benefit from the different motifs (Figure 7C). Among human pri-miRNAs, these motifs were nonetheless highly enriched, with 79% of the conserved human miRNAs containing at least one of the three motifs (Figure 7D).

The motifs were not enriched in *C. elegans* pri-miRNAs (Figures 7E) and, when added to the *C. elegans* pri-miRNAs, conferred more efficient processing in mammalian cells (Figure 7B, S7G, and S7H). These experiments also showed the benefit of disrupting pairing normally present at positions −14 and +12 of the *C. elegans* miRNAs. The presence of pairing that is inhibitory to mammalian processing suggests that measurement from the base of the helix might also differ

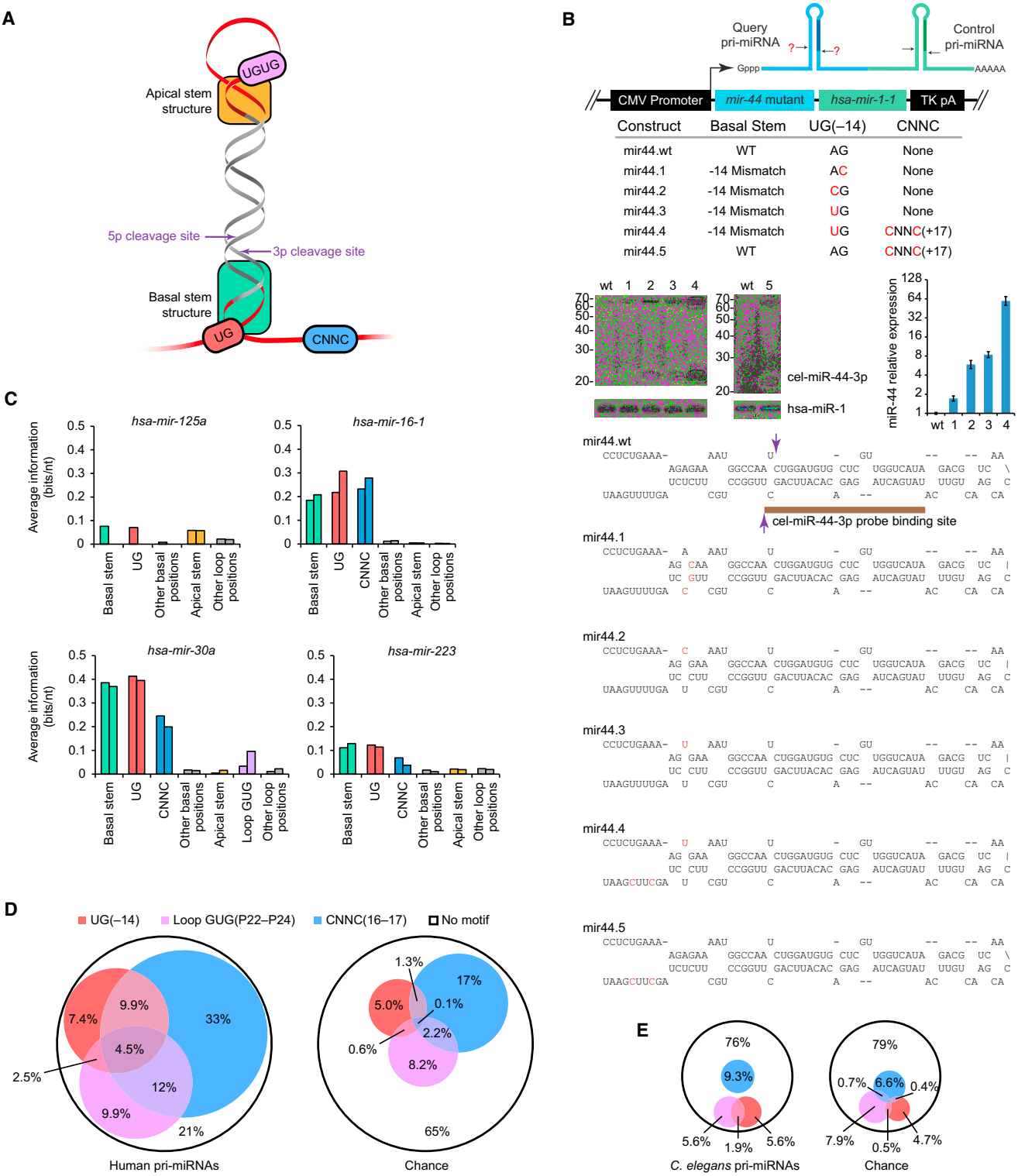


Figure 7. Structural and Primary-Sequence Features Important for Human Pri-miRNA Processing

(A) Summary of human pri-miRNA determinants identified or confirmed in this study.

(B) Processing enhancement from adding human pri-miRNA features to *C. elegans mir-44*. Changes that introduced the listed features were incorporated into *mir-44* within the bicistronic expression vector (top). Secondary structures are shown for mutations predicted to affect the wild-type basal stem (bottom; Drosha cleavage sites, purple arrowheads). After transfection into HEK293T cells, accumulation of miR-44-3p was assessed on RNA blots (middle), with the graph plotting increased miR-44-3p expression normalized to that of the hsa-miR-1 control (geometric mean ± standard error, n = 3). Adding a CNNC to the

(legend continued on next page)

in nematodes. Thus, despite the many broadly conserved features of miRNAs, some primary-sequence features and some secondary-structure features differ in mammals and nematodes.

About a fifth of human pri-miRNAs lack all three newly identified primary-sequence determinants (Figure 7D). These are attractive subjects for further study, in that the approach implemented here presumably would identify additional unique determinants used by these pri-miRNAs. Other determinants probably also exist at the Microprocessor cleavage site and nearby stem regions, which were inaccessible to our approach as implemented. Indeed, point mutations that disrupt pairing in the middle of the stem dramatically impair processing (Gottwein et al., 2006; Duan et al., 2007; Jazdzewski et al., 2008; Sun et al., 2009), and the SR-domain splicing factor SF2/ASF is reported to enhance the processing of *mir-7-1* by binding a motif in the stem near the cleavage site (Wu et al., 2010). Hinting at the possibility of additional primary-sequence preferences within the stem are results from both bacterial RNase III and fungal homologs (Rnt1 and Pac1), which prefer specific base-pair identities near the cleavage site (Lamontagne and Elela, 2004).

The emerging picture is that pri-miRNA recognition is a modular phenomenon in which each module contributes modestly, and each pri-miRNA depends on individual modules to varying degrees. Our results quantify the relative importance of each known module for each pri-miRNA (Figure 7C). Pairing within the basal stem was crucial, as expected (Lim et al., 2003b; Han et al., 2006). In addition, all four miRNAs made use of the basal UG motif, which provided information content per nucleotide resembling that provided by the basal-stem nucleotides. For the three miRNAs that used the CNNC SRp20-binding site, its importance was also comparable to that of the basal stem nucleotides. Compared to the nucleotides within these motifs, other flanking nucleotides contributed very little.

Apical and terminal loop elements were less important than the basal motifs (Figure 7C). We detected significant contributions only in *mir-125a*, in which the apical stem nucleotides were as important as the basal stem nucleotides, and in *mir-30a*, in which the loop UGUG motif contributed some information, albeit less than any of the three other features. Together, the features described here explained 61%–78% of the information content in the selected sequences. The remaining information content was diffusely distributed among the other partially randomized positions and might have mostly reflected avoidance of detrimental alternative structures.

Knowledge of biogenesis features will aid in interpreting human mutations. For example, reduced miR-16 expression associated with chronic lymphocytic leukemia (CLL) is typically due to deletions spanning the intron containing *mir-15a* and *mir-16-1* (Calin et al., 2002). However, 2 of 75 CLL patients studied had tumors that retain the pri-miRNA hairpins and instead carried a germline C > T single-nucleotide polymorphism (SNP) downstream of the *mir-16-1* hairpin (Calin et al., 2005). This SNP lowers overexpression of miR-16 in HEK293 cells, and in both patients heterozygosity for the SNP was lost in the leukemic cells (Calin et al., 2005). This SNP corresponds to the first C in the *mir-16-1* CNNC, which explains why it lowers miR-16 accumulation and leads to CLL: it affects pri-miRNA processing by disrupting SRp20 recruitment. Discovery of additional features for pri-miRNA recognition and processing might lead to improved diagnostic and therapeutic tools in cancer and other diseases in which miRNAs are dysregulated.

EXPERIMENTAL PROCEDURES

Ectopic Pri-miRNA Expression

Plasmids were derived from pcDNA3.2/V5-DEST and pMT-DEST (Invitrogen) for expression in HEK293 and S2 cells, respectively. Query pri-miRNA sequences and the human *pri-mir-1-1* sequence were cloned such that the query pri-miRNAs were transcriptionally fused upstream of *mir-1-1*. HEK293 and S2 cells were transfected using Lipofectamine 2000 and Cellfectin (Invitrogen), respectively. After 36–48 hr, total RNA was extracted, and miRNA expression was assayed by RNA blots, ribonuclease protection assays (Invitrogen), and high-throughput sequencing (Chiang et al., 2010). For additional details including the data analysis pipeline, see Extended Experimental Procedures.

Binding and Cleavage Assays

To assay binding, we radiolabeled and mixed T7-transcribed competitor and reference pri-miRNA substrates in an equimolar ratio, then incubated them with limiting amounts of immunopurified catalytically impaired Microprocessor (Lee and Kim, 2007; Han et al., 2009). RNA-protein complexes were filtered on Immobilon-NC nitrocellulose discs (Whatman), and RNA extracted from the filter was resolved on 5% polyacrylamide gels. To assay cleavage, we incubated labeled substrates with Microprocessor lysate, which was prepared from cells overexpressing Drosha and DGCR8 (Lee and Kim, 2007). After extraction using Tri-Reagent (Ambion), substrates and products were resolved on denaturing 5% polyacrylamide gels. For additional details, see Extended Experimental Procedures.

Synthesis and Selection of Pri-miRNA Variants

Templates for T7 transcription were assembled from oligonucleotides (IDT) synthesized using nucleoside phosphoramidite mixtures designed to introduce variability at specified positions (Table S1). Sequences encoding the HDV self-cleaving ribozyme were appended so that ribozyme cleavage would

wild-type sequence (construct mir44.5) enhanced processing ≥ 20-fold (geometric mean of triplicate experiment), a lower bound set by the wild-type background.

(C) Contributions of individual features to in vitro processing measured as average information content per nucleotide. If available, results from two time points are shown.

(D) Enrichment of primary-sequence motifs in human pri-miRNAs conserved to mouse (Table S2). Pri-miRNAs were classified based on whether they had the basal UG, the apical GUG or UGU, or the downstream CNNC motif (left). Expectations by chance (right) were estimated based on the nucleotide composition of upstream, pre-miRNA, and downstream regions of human pri-miRNAs for the basal UG, apical GUG or UGU, and CNNC motifs, respectively.

(E) A search for human motifs in *C. elegans* pri-miRNAs (Table S2). Pri-miRNAs were analyzed as in (D); the smaller diagrams reflect the smaller number of analyzed pri-miRNAs.

See also Figure S7.

generate transcripts with defined 3′ ends. Template pools were transcribed using T7 RNA polymerase, and after treatment with TurboDNase (Ambion) RNA was purified on denaturing polyacrylamide gels. After dephosphorylation of 5′ and 3′ ends using calf intestinal phosphatase (NEB) and T4 polynucleotide kinase (T4 PNK, NEB), followed by 5′ phosphorylation using T4 PNK, transcripts were circularized using T4 RNA ligase 1 (NEB) and gel purified. RNA pools were incubated with Microprocessor lysate, and after gel purification, cleavage products were ligated to oligonucleotide adaptors, reverse transcribed, amplified, and Illumina sequenced (75 nt paired-end reads). In parallel, the initial pool of RNA was also reverse transcribed, amplified, and sequenced. Selections for examining binding or apical stem-loops were similar, except transcripts were not circularized. For additional details including the data analysis pipeline, see Extended Experimental Procedures.

Motif Enrichment

Enrichment of a motif within pri-miRNAs of a species was evaluated by comparing to 100,000 cohorts of miRNAs in which the upstream, downstream, and pre-miRNA sequences were independently shuffled, preserving dinucleotide frequencies. The numbers of miRNAs that contained a match to the motif in the actual and shuffled cohorts were used to compute an empirical p value. A list of the representative pri-miRNAs used for analyses is provided (Table S2). For additional details, see Extended Experimental Procedures.

Site-Specific Crosslinking

The *mir-30a* pri-miRNA crosslinking substrate was assembled using T4 RNA ligase 2 (NEB) and a DNA splint to join an in vitro-transcribed 5′ fragment to a synthetic 3′ fragment containing a 3′-terminal biotin and a 4-thiouridine within the CNNC motif (Dharmacon). This crosslinking substrate was incubated in Microprocessor lysate and exposed to 1000 mJ of 365 nm UV light in a Stratalinker (Stratagene). For purification of RNA-protein complexes for mass spectrometry, complexes were captured on streptavidin-coated magnetic beads (Invitrogen), washed, and eluted with RNase T1 (Ambion), which cleaves after G. Eluted complexes either were separated on SDS gels and analyzed by HPLC/tandem mass spectrometry or were immunoprecipitated and analyzed by SDS gel. For additional details, see Extended Experimental Procedures.

ACCESSION NUMBERS

The Short Read Archive accession number for the sequencing data reported in this paper is SRA051323.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, two tables, and Extended Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.01.031.

ACKNOWLEDGMENTS

We thank D. Shechner, C. Jan, O. Rissland, D. Weinberg, J. Ruby, J. Nam, and V.N. Kim for valuable discussions; O. Rissland for comments on this manuscript; L. Schoenfeld and J. Lassar for technical assistance; J. Stévenin for 9G8 antibody; V.N. Kim and T. Tuschl for plasmids; the Whitehead Institute Genome Technology Core for sequencing; and E. Spooner for mass spectrometry. This work was supported by NIH grants GM067031 and T32GM007753. D.B. is an Investigator of the Howard Hughes Medical Institute.

Received: April 10, 2012
Revised: October 28, 2012
Accepted: January 14, 2013
Published: February 14, 2013

REFERENCES

Änkö, M.L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K.M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol. *13*, R17.

Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev. *22*, 2773–2785.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Bedard, K.M., Daijogo, S., and Semler, B.L. (2007). A nucleo-cytoplasmic SR protein functions in viral IRES-mediated translation initiation. EMBO J. *26*, 459–467.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Genet. *37*, 766–770.

Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S., et al. (2006). Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res. *16*, 1289–1298.

Brummelkamp, T.R., Bernards, R., and Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. Science *296*, 550–553.

Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., et al. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc. Natl. Acad. Sci. USA *99*, 15524–15529.

Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M.V., Visone, R., Sever, N.I., Fabbri, M., et al. (2005). A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N. Engl. J. Med. *353*, 1793–1801.

Cavaloc, Y., Popielarz, M., Fuchs, J.P., Gattoni, R., and Stévenin, J. (1994). Characterization and cloning of the human splicing factor 9G8: a novel 35 kDa factor of the serine/arginine protein family. EMBO J. *13*, 2639–2649.

Cavaloc, Y., Bourgeois, C.F., Kister, L., and Stévenin, J. (1999). The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA *5*, 468–483.

Cheloufi, S., Dos Santos, C.O., Chong, M.M., and Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584–589.

Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. Science *303*, 83–86.

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. *24*, 992–1009.

Chung, W.J., Agius, P., Westholm, J.O., Chen, M., Okamura, K., Robine, N., Leslie, C.S., and Lai, E.C. (2011). Computational and experimental identification of mirtrons in *Drosophila melanogaster* and *Caenorhabditis elegans*. Genome Res. *21*, 286–300.

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., et al. (2010). A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science *328*, 1694–1698.

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

Duan, R., Pak, C., and Jin, P. (2007). Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. Hum. Mol. Genet. *16*, 1124–1131.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., and Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol. Cell *41*, 733–746.

Feng, Y., Zhang, X., Song, Q., Li, T., and Zeng, Y. (2011). Drosha processing controls the specificity and efficiency of global microRNA expression. Biochim. Biophys. Acta *1809*, 700–707.

Gottwein, E., Cai, X., and Cullen, B.R. (2006). A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. J. Virol. *80*, 5321–5326.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235–240.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. Cell *106*, 23–34.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H., and Kim, V.N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev. *18*, 3016–3027.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887–901.

Han, J., Pedersen, J.S., Kwon, S.C., Belair, C.D., Kim, Y.K., Yeom, K.H., Yang, W.Y., Haussler, D., Blelloch, R., and Kim, V.N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. Cell *136*, 75–84.

Hargous, Y., Hautbergue, G.M., Tintaru, A.M., Skrisovska, L., Golovanov, A.P., Stevenin, J., Lian, L.Y., Wilson, S.A., and Allain, F.H. (2006). Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. EMBO J. *25*, 5126–5137.

Heinrichs, V., and Baker, B.S. (1995). The *Drosophila* SR protein RBP1 contributes to the regulation of doublesex alternative splicing by recognizing RBP1 RNA target sequences. EMBO J. *14*, 3987–4000.

Hofacker, I.L., and Stadler, P.F. (2006). Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics *22*, 1172–1176.

Huang, Y., and Steitz, J.A. (2001). Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. Mol. Cell *7*, 899–905.

Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. Science *293*, 834–838.

Jazdzewski, K., Murray, E.L., Franssila, K., Jarzab, B., Schoenberg, D.R., and de la Chapelle, A. (2008). Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. Proc. Natl. Acad. Sci. USA *105*, 7269–7274.

Jia, R., Li, C., McCoy, J.P., Deng, C.X., and Zheng, Z.M. (2010). SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. Int. J. Biol. Sci. *6*, 806–826.

Jumaa, H., Wei, G., and Nielsen, P.J. (1999). Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. Curr. Biol. *9*, 899–902.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209–216.

Lamontagne, B., and Elela, S.A. (2004). Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. J. Biol. Chem. *279*, 2231–2241.

Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and its *D. melanogaster* homolog are required for miRNA biogenesis. Curr. Biol. *14*, 2162–2167.

Lee, Y., and Kim, V.N. (2007). In vitro and in vivo assays for the activity of Drosha complex. Methods Enzymol. *427*, 89–106.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415–419.

Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. (2003a). Vertebrate microRNA genes. Science *299*, 1540.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003b). The microRNAs of *Caenorhabditis elegans*. Genes Dev. *17*, 991–1008.

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

Macrae, I.J., Zhou, K., Li, F., Repic, A., Brooks, A.N., Cande, W.Z., Adams, P.D., and Doudna, J.A. (2006). Structural basis for double-stranded RNA processing by Dicer. Science *311*, 195–198.

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

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev. *16*, 948–958.

Pan, T., and Uhlenbeck, O.C. (1992). In vitro selection of RNAs that undergo autolytic cleavage with Pb2+. Biochemistry *31*, 3887–3895.

Park, J.E., Heo, I., Tian, Y., Simanshu, D.K., Chang, H., Jee, D., Patel, D.J., and Kim, V.N. (2011). Dicer recognizes the 5′ end of RNA for efficient and accurate processing. Nature *475*, 201–205.

Pitt, J.N., and Ferré-D'Amaré, A.R. (2010). Rapid construction of empirical RNA fitness landscapes. Science *330*, 376–379.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83–86.

Schaal, T.D., and Maniatis, T. (1999). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol. Cell. Biol. *19*, 1705–1719.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199–208.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., and Mann, R.S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D.A., Sommer, S.S., and Rossi, J.J. (2009). SNPs in human miRNA genes affect biogenesis and function. RNA *15*, 1640–1651.

Swartz, J.E., Bor, Y.C., Misawa, Y., Rekosh, D., and Hammarskjold, M.L. (2007). The shuttling SR protein 9G8 plays a role in translation of unspliced mRNA containing a constitutive transport element. J. Biol. Chem. *282*, 19844–19853.

Wilson, D.S., and Szostak, J.W. (1999). In vitro selection of functional nucleic acids. Annu. Rev. Biochem. *68*, 611–647.

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

Wyatt, J.R., Sontheimer, E.J., and Steitz, J.A. (1992). Site-specific cross-linking of mammalian U5 snRNP to the 5′ splice site before the first step of pre-mRNA splicing. Genes Dev. *6*(12B), 2542–2553.

Yeom, K.H., Lee, Y., Han, J., Suh, M.R., and Kim, V.N. (2006). Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. Nucleic Acids Res. *34*, 4622–4629.

Zahler, A.M., Neugebauer, K.M., Stolk, J.A., and Roth, M.B. (1993). Human SR proteins and isolation of a cDNA encoding SRp75. Mol. Cell. Biol. *13*, 4023–4028.

Zeng, Y., and Cullen, B.R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. J. Biol. Chem. *280*, 27595–27603.

Zeng, Y., Yi, R., and Cullen, B.R. (2005). Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. EMBO J. *24*, 138–148.

Zhang, H., Kolb, F.A., Jaskiewicz, L., Westhof, E., and Filipowicz, W. (2004). Single processing center models for human Dicer and bacterial RNase III. Cell *118*, 57–68.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res. *37*, e151.

Appendix B.

RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins

Nicole Lambert[1]*, Alex Robertson[1,2]*, Mohini Jangi[1], Sean McGeary[1,4], Phillip A. Sharp[1,3], and Christopher B. Burge[1,3]

[1]Department of Biology
[2]Program in Computational and System Biology
[3]Koch Institute for Integrative Cancer Research
[4]Whitehead Institute for Biomedical Research
Massachusetts Institute of Technology, Cambridge, MA 02142, USA
*These authors contributed equally to this work.

Published as:
Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–90.

Molecular Cell
Resource

RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins

Nicole Lambert,[1,5] Alex Robertson,[1,2,5] Mohini Jangi,[1] Sean McGeary,[1,4] Phillip A. Sharp,[1,3] and Christopher B. Burge[1,3,*]
[1]Department of Biology
[2]Program in Computational and Systems Biology
[3]Koch Institute for Integrative Cancer Research
[4]Whitehead Institute for Biomedical Research
Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[5]Co-first authors
*Correspondence: cburge@mit.edu
http://dx.doi.org/10.1016/j.molcel.2014.04.016

SUMMARY

Specific protein-RNA interactions guide posttranscriptional gene regulation. Here, we describe RNA Bind-n-Seq (RBNS), a method that comprehensively characterizes sequence and structural specificity of RNA binding proteins (RBPs), and its application to the developmental alternative splicing factors RBFOX2, CELF1/CUGBP1, and MBNL1. For each factor, we recovered both canonical motifs and additional near-optimal binding motifs. RNA secondary structure inhibits binding of RBFOX2 and CELF1, while MBNL1 favors unpaired Us but tolerates C/G pairing in motifs containing UGC and/or GCU. Dissociation constants calculated from RBNS data using a novel algorithm correlated highly with values measured by surface plasmon resonance. Motifs identified by RBNS were conserved, were bound and active in vivo, and distinguished the subset of motifs enriched by CLIP-Seq that had regulatory activity. Together, our data demonstrate that RBNS complements crosslinking-based methods and show that in vivo binding and activity of these splicing factors is driven largely by intrinsic RNA affinity.

INTRODUCTION

RNA binding proteins (RBPs) bind sequence and/or structural motifs in nuclear pre-mRNAs to direct their processing and bind mature mRNAs to control their translation, localization, and stability. RBPs of the Rbfox, CUG-BP/Elav-like (CELF), and muscleblind-like (MBNL) families are important and highly conserved regulators of developmental and tissue-specific alternative splicing.

Rbfox2, a close homolog of Rbfox1 (Underwood et al., 2005), is required for neural development (Gehman et al., 2012), regulates epithelial-mesenchymal transition (EMT) (Baraniak et al., 2006), and is required for human embryonic stem cell (ESC) survival (Yeo et al., 2009). The consensus binding motif for Rbfox proteins—UGCAUG or simply GCAUG—has been determined by systematic evolution of ligands by exponential enrichment (SELEX) and is conserved from nematodes through vertebrates (Jin et al., 2003; Ponthier et al., 2006). However, the iterative selection steps used in SELEX favor recovery of just the strongest binding motifs and may not detect moderate and lower affinity motifs. Only about one-third to one-half of Rbfox2 binding sites identified in vivo contain these canonical motifs (Jangi et al., 2014; Yeo et al., 2009), but it has remained unclear whether this RBP can recognize other sequence motifs. In general, motifs recognized by RBPs with lower affinity are more challenging to characterize, but such motifs may play biological roles that are as important as those played by higher affinity motifs. For RBPs that accumulate during development, like MBNLs, higher affinity motifs may be bound at earlier time points, while lower affinity motifs may specify regulation at later developmental time points or only in certain cell types where the RBP accumulates to high levels.

CELF1 and MBNL1 proteins are functionally linked by their roles in development and disease, often regulating the same splicing targets in an antagonistic fashion. In heart development, during which CELF protein levels decrease and MBNL proteins accumulate, this antagonism may sharpen developmental splicing transitions (Kalsotra et al., 2008). This developmental expression pattern reverses that seen in the muscle wasting disease myotonic dystrophy type 1 (DM1), in which expanded CUG repeats in the 3′ UTR of *DMPK* mRNAs reduce available cellular levels of MBNL proteins by sequestration (Mankodi et al., 2005; Taneja et al., 1995), and CELF1 proteins are stabilized by hyperphosphorylation (Kuyumcu-Martinez et al., 2007). CELF1 has three RNA recognition motifs (RRMs) that bind motifs with consensus UGU (Ladd et al., 2001; Marquis et al., 2006). MBNL1 has two pairs of zinc fingers that are reported to bind preferentially to YGCY (Y = C or U) motifs (Ho et al., 2004). To date, it has remained unclear whether MBNL1 primarily recognizes single- or double-stranded RNA elements. CUG repeat RNA crystallizes as an A-form helix (Mooers et al., 2005), with C and G bases paired and Us unpaired, and additional biochemical studies have shown that a mismatched RNA hairpin structure is important for recognition by MBNL1 (Warf and Berglund,

2007). However, structures of MBNL1 zinc fingers cocrystallized with CGCUGU RNA suggested that MBNL1 recognizes single-stranded RNA (Teplova and Patel, 2008). Additionally, the roles of motif spacing and of intervening sequences between tandem motifs remain largely uncharacterized.

Widely used methods for mapping protein-RNA interactions in vivo based on ultraviolet cross-linking and immunoprecipitation (CLIP) (Ule et al., 2003; Underwood et al., 2005) have contributed to understanding of posttranscriptional regulation. However, these techniques are laborious and require many selection steps that likely introduce various types of bias. Motif analysis from CLIP data is complicated by the fact that it does not distinguish binding by a single protein from binding of a protein complex, and it may preferentially detect uridine-rich sequences (Sugimoto et al., 2012). Iterative binding approaches like SELEX, including recent high-throughput versions (Campbell et al., 2012), identify consensus motifs but are not quantitative and are biased toward the highest affinity motifs. A newer method, RNAcompete, uses in vitro RNA-protein binding followed by microarray analysis, enabling high-throughput identification of RNA binding motifs (Ray et al., 2009, 2013). However, the number of probes assayed and the low temperatures typically used make it difficult to analyze effects of RNA secondary structure on RNA binding, and RNAcompete does not yield K_d values. Quantitative biophysical measurements including K_d values can be obtained from methods such as electrophoretic mobility shift assays (EMSAs) or surface plasmon resonance (SPR), but their throughput is quite low.

To better characterize the functions of biologically important RBPs, we sought to develop a method that would measure affinities to the full spectrum of bound RNAs in a quantitative and high-throughput manner. Methods for characterizing protein/DNA interactions that are both high-throughput and quantitative have been developed, including HT-SELEX and Bind-n-Seq, both of which use one-step binding to a pool of randomized DNA in vitro followed by deep sequencing (Jolma et al., 2010; Zykovich et al., 2009), and HiTS-FLIP, which directly measures protein bound to double-stranded DNA on a flow cell (Nutiu et al., 2011). We adapted the general approach used by HT-SELEX and Bind-n-Seq to the study of protein-RNA interactions in vitro in a method we call "RNA Bind-n-Seq" (RBNS). Our method adapts and extends these protein/DNA interaction assays in two important ways. First, we use multiple RBP concentrations to optimize analysis at different ranges of affinity. Second, we have expanded the analytical framework to more accurately estimate relative dissociation constants and to assess the effects of RNA secondary structure on binding. RBNS analyses of RBFOX2, CELF1, and MBNL1 yielded comprehensive portraits of the sequence and RNA secondary structural determinants of RNA recognition by these factors. Analysis of data from systems in which these RBPs were depleted or inducibly overexpressed in mouse cells provided evidence of function for both noncanonical and canonical binding motifs identified in vitro. We observed good correlation between in vitro and in vivo binding overall, but we found that motifs enriched by CLIP only (but not by RBNS) are not associated with regulatory activity. Therefore, RBNS aids in identification of high-confidence splicing-associated binding sites and is complementary to CLIP.

RESULTS

Design Considerations for RBNS Experiments

RBNS is designed to dissect the sequence and RNA structural preferences of RBPs. A recombinantly expressed and purified RBP is incubated with a pool of randomized RNAs at several different protein concentrations, typically ranging from low nanomolar to low micromolar (Figure 1A). The RNA pool typically consists of random RNAs of length $\lambda = 40$ nt flanked by short primers used to add the adapters needed for deep sequencing. This RNA pool design simplifies library preparation, avoids biases that can result from RNA ligation, and ensures that any bacterial RNA carried over from protein expression will not contaminate the sequenced library. (In the unusual case where the RBP has significant affinity to primer RNA, different primer sequences must be substituted.) In each experiment, the RBP is captured via a streptavidin binding peptide (SBP) tag. RBP-bound RNA is reverse-transcribed into cDNA, and barcoded sequencing adapters are added by PCR to produce libraries for deep sequencing. Libraries corresponding to the input RNA pool and to five or more RBP concentrations (including zero RBP concentration as an additional control) are sequenced in a single Illumina HiSeq 2000 lane, typically yielding at least 15–20 million reads per library.

Most RBPs bind single-stranded RNA sequence motifs 3–8 nt in length (Stefl et al., 2005). Here, we performed one experiment using the RBFOX2 RRM with short oligonucleotides ($\lambda = 10$ nt). However, we soon realized that use of longer sequences ($\lambda = 40$ nt) provided comparable affinity measurements to short, linear motifs of size *k* (*k*mers) in the range of interest (about 3–10 nt; Figure S1 available online) while also enabling assessment of RNA secondary structural and other contextual effects on binding that cannot be assessed using 10mers. Size $\lambda = 40$ nt is closer to the in vivo situation where RBPs typically bind long RNAs, but it is within the range where structure can be most accurately predicted by thermodynamic RNA folding algorithms (Hofacker, 2003).

RBNS Comprehensively Identifies Known and Secondary Motifs of RBPs

RBNS was performed using recombinant RBFOX2, MBNL1, and CELF1 proteins. For each protein, at each of several concentrations, motif read enrichment ("R") values were calculated for each *k*mer (for $k = 5$, 6, and 7) as the ratio of the frequency of the *k*mer in the selected pool to the frequency in the input RNA library. In our typical zero concentration experiment, 99.9% of 6mers had R values less than 1.19, and the highest value was 1.21, indicating little if any sequence bias from the apparatus. The false discovery rate (FDR) was 1.2% for CELF1 7mers, as judged by the 0 nM RBP experiment, and was ~0% for the other proteins (Supplemental Experimental Procedures).

For RBFOX2, at all concentrations $\geq$14 nM the 6mer UGCAUG had the highest R value (Figure 1B and below), confirming this well-known motif as the highest affinity 6mer. The enrichment of UGCAUG reached a maximum R of 22 at a protein concentration of 365 nM (Figure 1B). We derived an equation relating the observed R value to the relative affinity (ratio of dissociation constants, B) between nonspecific and specific binding

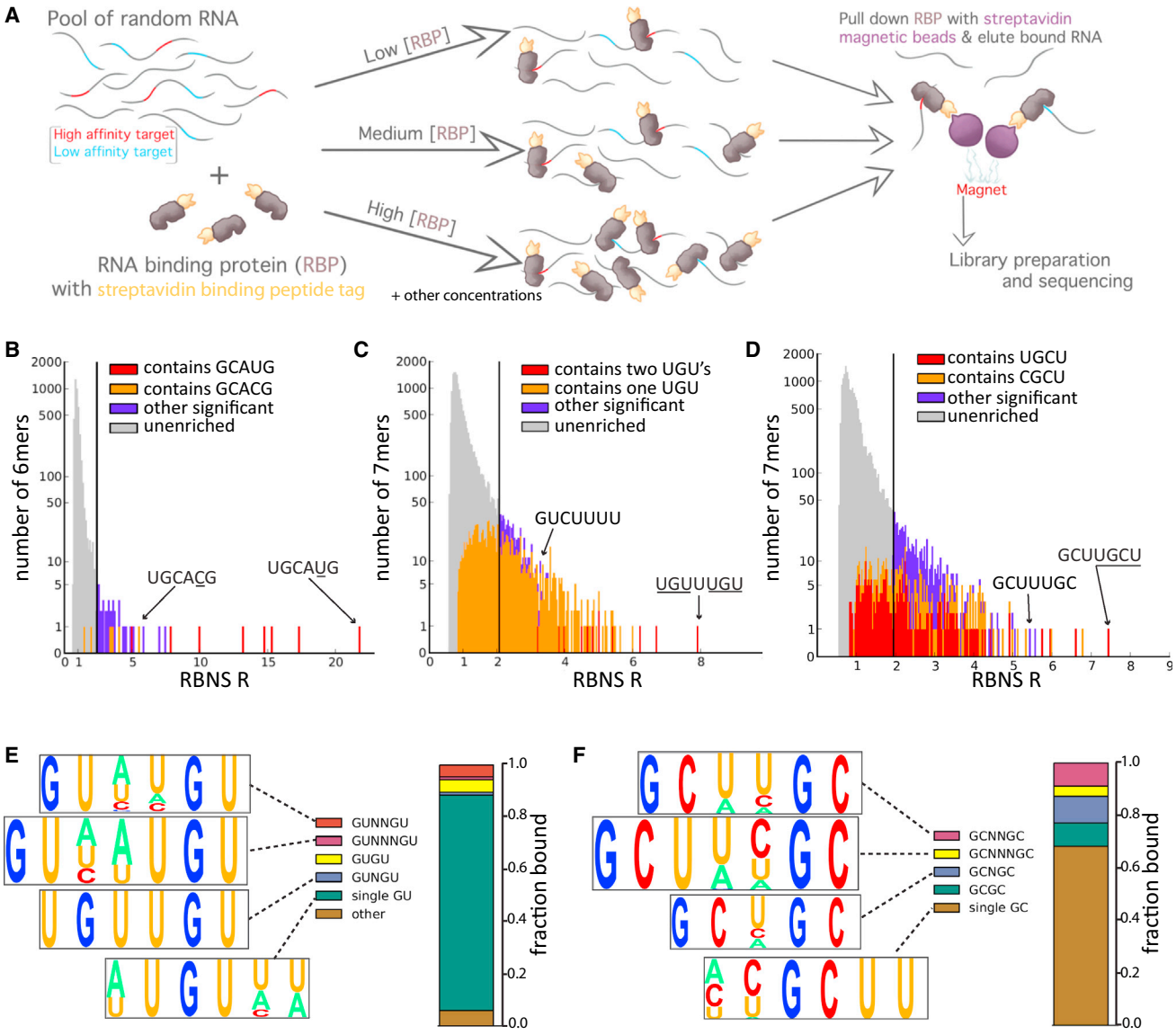


Figure 1. RNA Bind-n-Seq overview and Motif Enrichment Analysis

(A) Overview of the experimental method. Tagged protein is incubated with a diverse pool of RNA oligonucleotides of fixed concentration at each of several concentrations of protein. The RBP is pulled down using streptavidin-coated magnetic beads, and the associated RNA is sequenced. The counts of sequences in this library are used to estimate proportions of bound RNA molecules, in comparison to input RNA, which is also sequenced.

(B) Stacked histogram showing the distribution of RBNS R values of all RNA 6mers in the RBFOX2 experiment at a protein concentration of 365 nM. 6mers that contain specific 5mers, whether significant or not are shown in red or orange; other 6mers are colored based on whether their R value is at least 2 SD above the mean (purple) or not (gray). A log scale is used for the y axis.

(C) As in (B), but shows distribution of R values for all 7mers for CELF1 at a protein concentration of 64 nM.

(D) As in (C), but shows distribution of all 7mers for MBNL1 at a concentration of 250 nM.

(E) Visualization of CELF1 binding preferences. The sequence content (displayed as a pictogram with letter height proportional to frequency) and estimated bound fraction of four groups of 7mer motifs are shown. The top 50 7mers were grouped and aligned based on their content and spacing of GU submotifs (Figure S1D).

(F) Visualization of the Mbnl1 binding preferences. As in (E), but based on the top 50 7mer motifs for MBNL1, grouped by spacing of GC submotifs (alignments shown in Figure S1E).

See also Figure S1.

under idealized conditions (Supplemental Experimental Procedures, Equation 18, "RB Equation"). With R = 22, k = 6, and λ = 40, this equation implies at least ~900-fold higher binding affinity to UGCAUG than to nonspecific 6mers. All eight of the 6mers that contain GCAUG had significant R values (Figure 1B), consistent with the known affinity of Rbfox proteins for this 5mer (Jin et al., 2003). Several 6mers containing GCACG were also significant, indicating that this 5mer represents an alternate

RBFOX2 binding motif. Certain other 6mers not containing GCAUG or GCACG, but often containing GCAU, also had significant R values, suggesting that RBFOX2 has some affinity for other RNA motifs as well (Table S1).

Proteins of the CELF family preferentially bind to UG- and UGU-containing motifs (Marquis et al., 2006; Timchenko et al., 1996). For CELF1, a large number of 6mer and 7mer motifs had significant R values (7mer analysis shown in Figure 1C). Inspection of these motifs showed that the highest R values were observed for 7mers containing two UGU triplets. In fact, all 7mers containing two UGUs were significantly enriched, suggesting that presence of two UGUs is sufficient for strong binding and that CELF1 tolerates presence or absence of a 1 nt spacer between UGUs (Figure 1C). The highest 7mer R value observed for CELF1, R = ~8 for UGUUUGU, implies >~250-fold binding affinity over background (RB Equation), somewhat below that of RBFOX2 for UGCAUG. This observation and the fatter tail of the R value distribution emphasize that CELF1 binds a broader spectrum of motifs with lower affinity than RBFOX2. Of the top fifty 7mers, all contained at least one UGU. However, not every motif containing a single UGU was significant, and some 7mers lacking UGU were significantly enriched, indicating that RNA recognition by CELF1 is complex. Inspection of the top fifty CELF1 7mers (Figure S1D) suggested that they can be clustered into four classes matching GUN_xGU for x = 0, 1, 2, and 3 and a fifth class containing a single GU (Figure 1E).

MBNL1 is known to favor binding to YGCY motifs in vitro by SELEX (Goers et al., 2010), and GCUU and UGCU were the top 4mers by CLIP-Seq (Wang et al., 2012). The most enriched 7mers for MBNL1 contained either YGCU or GCUU, often supplemented by a second GC. The most enriched 7mer, GCUUGCU, contained both of these 4mers and had an R value near 9, slightly higher than the top value for CELF1 (Figure 1D). Overall, 54% of 7mers containing YGCU, and 61% of those containing GCUU, but only 9% of those containing YGCC, had significant R values, suggesting that MBNL's specificity is better summarized as YGCU + GCUU rather than YGCY. MBNL1 7mers could be grouped into four classes matching GCN_xGC for x = 0, 1, 2, and 3 and a fifth class matching YGCU (Figure 1F). MBNL1's observed preference for multiple GCs with variable spacing is consistent with previous studies (Cass et al., 2011).

Relative Dissociation Constants Are Accurately Estimated from RBNS

To better understand the dependence of R values on RBP concentration and to assess the extent and effects of experimental noise, we modeled RBNS experiments and predicted the output under various assumptions. In an idealized setting in which an RBP binds a high-affinity motif X with K_d = 5 nM and several moderate affinity motifs Y each with K_d = 30 nM (assuming binding with 1:1 stoichiometry and a Hill coefficient of 1), the fraction of each motif bound is expected to follow essentially a sigmoidal function of RBP concentration, with half maximal binding to the motif occurring at a free protein concentration near the K_d value (Figure 2A). From the predicted binding fraction, assuming complete recovery of protein, the expected R value at each concentration can be determined under various assumptions about the affinity of the protein for nonspecific RNA and the amount of nonspecific RNA bound to the apparatus.

The modeled enrichment profiles (Figure 2B) show that R values of high-affinity motifs decrease as RBP concentrations become very high under all conditions tested. This effect is readily understood by considering that high RBP concentrations will tend to drive binding toward lower affinity RNAs (and high-affinity motifs may become saturated), resulting in a lower fraction of high-affinity motifs in RBP-bound RNA. These simulations also showed that even a small amount of nonspecific binding to the apparatus greatly reduces R values at very low RBP concentrations, because nonspecifically recovered RNA dilutes the small amount of specifically bound RNA. Together, these two effects produce a characteristic unimodal curve that peaks at intermediate RBP concentrations under a wide range of assumptions about affinities (Figure 2C).

Unimodal enrichment profiles for highly enriched *k*mers were observed for RBFOX2, CELF1, and MBNL1, in general agreement with our model under the assumption of moderate levels of nonspecific background (Figure 2D). In all cases, R values near 1 were observed at RBP concentrations of 0 nM and began to climb above 1 in the low (4–40) nM range, decreasing to near 1 at the highest (micromolar) protein concentrations. For each factor, the relative rankings of *k*mers obtained at different protein concentrations were highly correlated, supporting the assay's robustness (Table S2).

Next, we sought to estimate K_d values from RBNS data. The initial quantity of each *k*mer present was estimated based on the input RNA concentration (1 μM), and the concentration of bound RNA was then calculated from the total concentration of protein-RNA complex, measured by Bioanalyzer analysis (Experimental Procedures). The fraction of bound RNA attributable to binding at each specific *k*mer was then estimated using a novel "streaming *k*mer assignment" (SKA) algorithm (Supplemental Experimental Procedures). SKA generalizes the analytical approach of the RB equation in that it accounts for arbitrarily complex combinations of affinities to different *k*mers. The SKA algorithm assigns binding to a specific *k*mer in each sequence probabilistically, based on continually updated estimates of relative binding preferences, using multiple passes through the sequence read data (Supplemental Experimental Procedures), somewhat analogous to the streaming assignment of ambiguously mapping sequence reads to a genome introduced in the recently described eXpress algorithm (Roberts and Pachter, 2013). Using simulated read data, we observed that assignments of binding locations within reads are more accurate when using SKA than when using raw R values or B values inferred using the RB equation. In particular, SKA can distinguish bound motifs from motifs enriched through frequent overlap with bound motifs. For example, binding of RBFOX2 to GCAUG motifs will cause overlapping motifs of the form CAUGN (N = A, C, G, or T) to be enriched in bound reads even if these motifs have no affinity for RBFOX2 except when preceded by a G. In these cases, the degree to which the bound motif is preferentially enriched enables the SKA

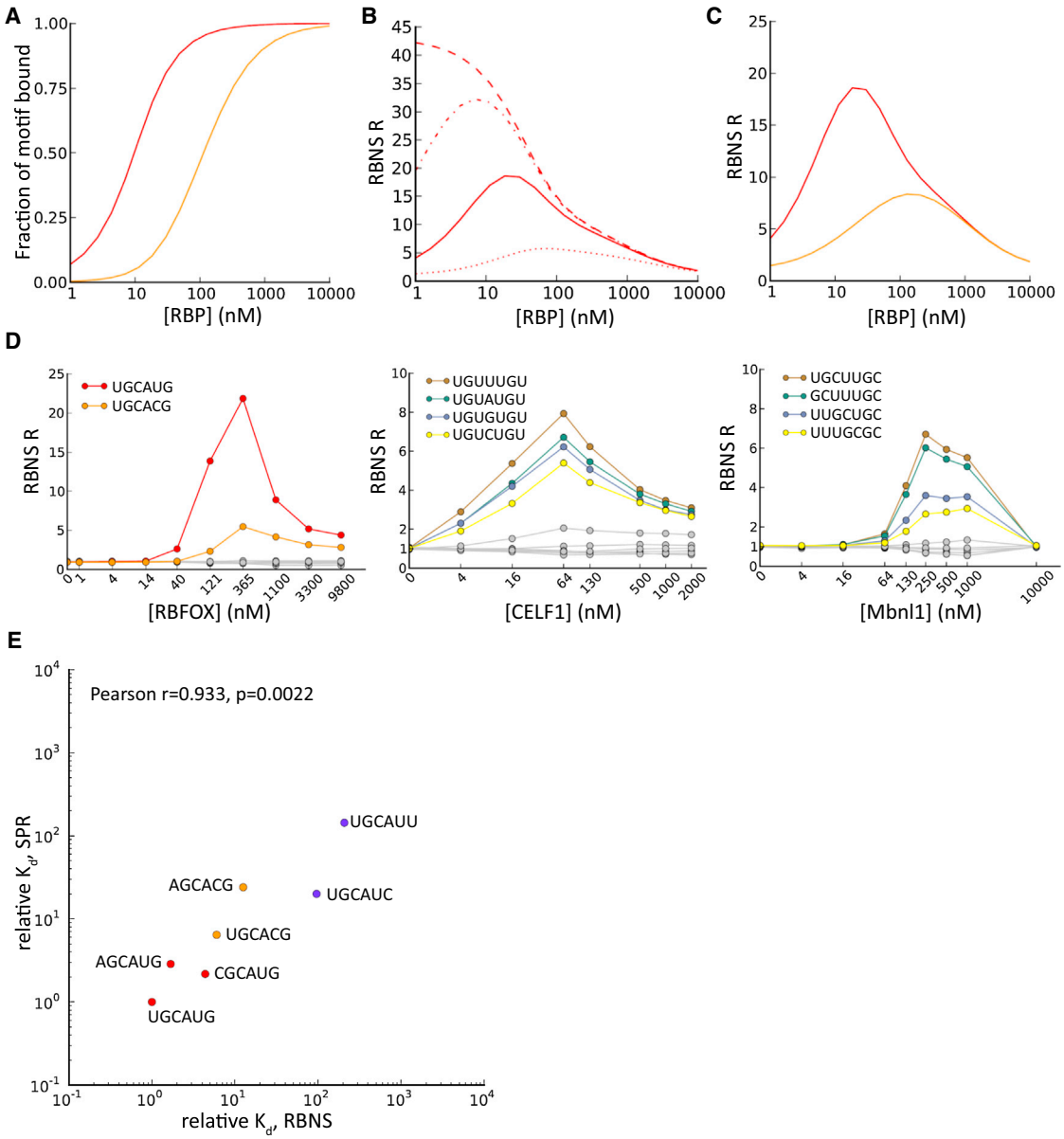


Figure 2. Modeling of RBNS Data and Estimation of Dissociation Constants

(A) Simulated output of RBNS under basic assumptions. Standard binding curves for two motifs of different binding affinities (see Results).

(B) Simulated RBNS R values for a single high-affinity 6mer motif as a function of protein concentration, under the assumption of different fixed amounts of nonspecific background (NSB) RNA recovery, independent of protein concentration (dashed no NSB; dash/dot: low NSB; solid: moderate NSB; dotted: high NSB).

(C) Simulated RBNS R values assuming presence of a single strong motif (red) and ten weaker motifs (orange), including moderate background nonspecific binding.

(D) RBNS R values for several top enriched 6mers or 7mers (colored) and several random 6mers/7mers (gray) are shown as a function of RBP concentration for each RBP studied. For RBFOX2, canonical UGCAUG and noncanonical UGCACG 6mers are shown. For CELF1, the four 7mers matching UGUNUGU are shown. For MBNL1, 7mers with two GCs at different spacings are shown, with flanking/intervening Us.

(E) Comparison of relative K_d values for several RBFOX2 6mers as estimated by RBNS (at RBFOX concentration 121 nM) and as measured by SPR. Correlation is significant by Pearson test ($R = 0.933$, $P = 2 \times 10^{-3}$). Motifs are colored as in Figure 1B.

See also Figure S4.

algorithm to effectively "learn" to assign lower probabilities (typically near background levels) to overlapping motifs (Figures S2 and S3).

Using estimates of bound and free *k*mer concentrations, we define the "relative" K_d value of a *k*mer as the ratio of the *k*mer's absolute dissociation constant to that of the highest affinity

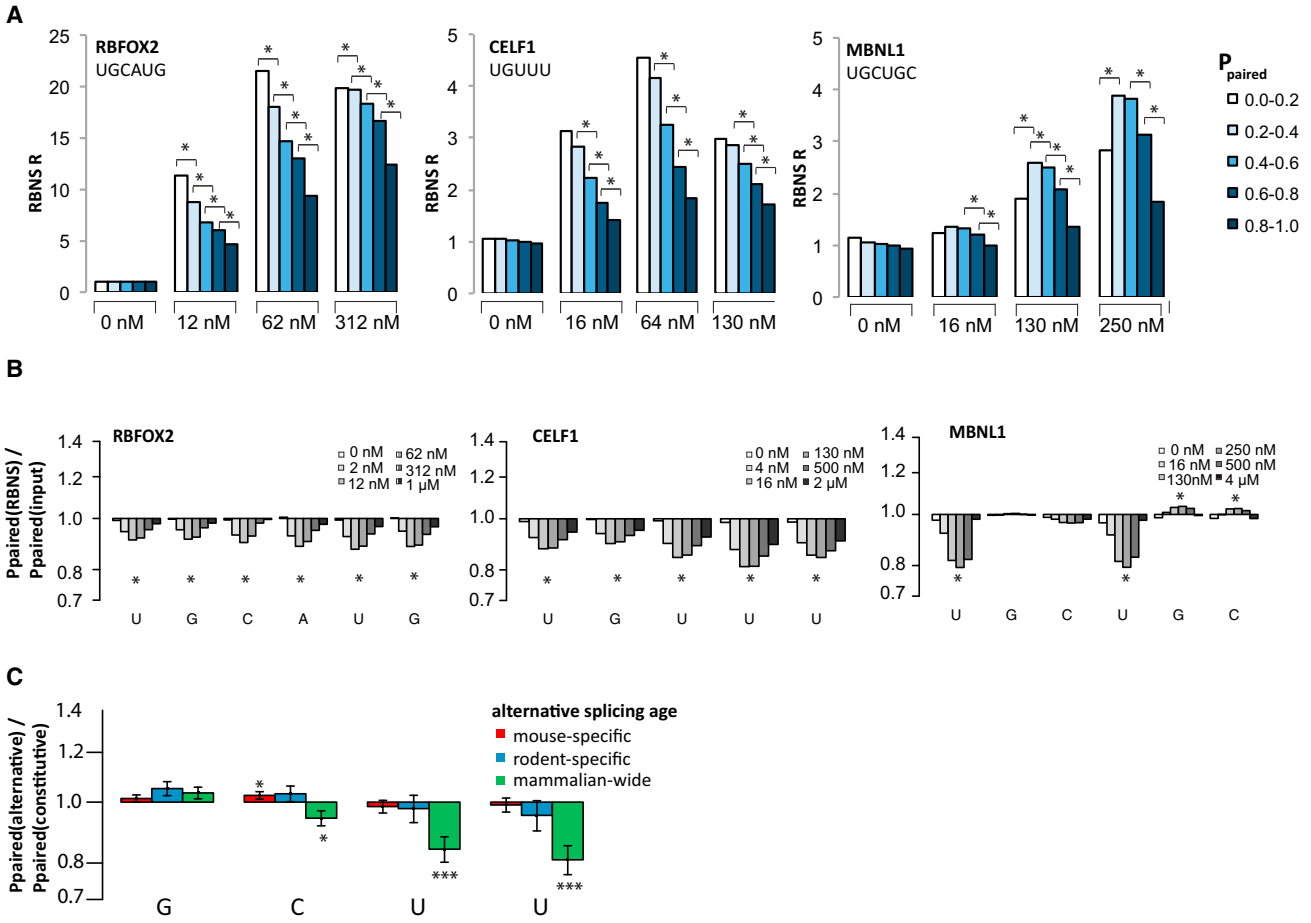


Figure 3. Impact of RNA Secondary Structure on Recognition of RNA Sequence Motifs

(A) Using rnafold, the average P_{paired} value across the bases in each instance of the indicated motif was used to assign each motif occurrence to one of the five P_{paired} bins indicated, and an R value was calculated at each RBP concentration for each bin as the frequency in the selected library divided by that in the input library. R values are shown for several concentrations of the three proteins, with asterisks indicating statistical significance (Z score > 2; $p < 0.05$) between adjacent structure bins (Experimental Procedures).

(B) The ratio of the mean value of P_{paired} in the bound library to that in the input control library is plotted on a log scale. Z scores were calculated for each selected library. Asterisks indicate bases where every selected library had |Z score| > 2 ($p < 0.05$).

(C) As in (B) for GCUU motifs located within 130 bases downstream of alternative exons of different evolutionary ages normalized to GCUU motifs in introns downstream of constitutive exons (Merkin et al., 2012). Error bars show SEM, and asterisks indicate significance by Wilcoxon rank-sum test (*$p < 0.05$; ***$p < 0.001$).

*k*mer. The *k*mers for which SKA predicts binding (those with absolute K_d < ~2000 nM) have relative K_d estimates spanning several orders of magnitude that are highly correlated to SPR measurements ($r = 0.94$, $p < 0.001$) (Figure 2E). Similarly high correlations were observed relative to previously measured SPR data for RBFOX1, a close paralog of RBFOX2 with identical RNA binding domain (Figure S4). Together, these observations demonstrate that RBNS yields quantitative measures of protein-RNA affinity.

Secondary Structure Inhibits Binding of Rbfox and CELF Proteins to RNA

RBNS can also be used to detect effects of RNA structure on binding of RBPs. We applied the thermodynamically based Vienna RNAfold algorithm (Hofacker, 2003) to sequence reads in order to assess the contribution of RNA structure to RBP:RNA interactions. In a motif-centric analysis, we analyzed folding of all RNAs harboring high-affinity UGCAUG, UGUUU, or UGCUGC motifs in RBFOX2, CELF1, or MBNL1 RBNS data sets, respectively (as well as other motifs), and in control libraries. The probability of intramolecular base pairing at each base in the motif was calculated from the energy-weighted ensemble of structures and averaged across the bases in the motif to give the "average base-pairing probability" (ABP). Sequence reads were then binned by their ABP, and R values were calculated separately for each combination of motif, protein concentration, and ABP bin. In these analyses, the bin with lowest ABP (0.0–0.2) was invariably the most enriched for both RBFOX2 and CELF1 binding at all nonzero RBP concentrations (Figure 3A), and R values decreased as ABP increased. Similar results were

obtained when analyzing other top motifs for these two factors. Together, these data suggest that RBFOX2 and CELF1 preferentially recognize single-stranded RNA motifs and that intramolecular base pairing directly competes with RBP recognition of these RNA motifs to a roughly similar extent for both proteins (Auweter et al., 2006; Edwards et al., 2013).

MBNL1 Binding Tolerates Pairing of GCs but Favors Unpaired Us

The RNA structure analysis for MBNL1 yielded a different pattern, with the highest R values observed for motifs with moderate ABP in the range 0.2–0.6. To better understand the impact of RNA structure on MBNL1 binding, we calculated the base-pairing probability for each base in bound sequences containing UGCUGC, and normalized to that of UGCUGC-containing RNAs in the input library, matching for C+G% content. This analysis showed no preference for lower base-pairing probabilities at GC positions but showed substantially reduced base pairing of Us in bound sequences (Figure 3B). A similar tolerance for pairing of the central GC dinucleotide and preference for unstructured flanking pyrimidine bases was observed for all high-affinity MBNL1 motifs tested, including UGCUU, GCUUGC, CGCUU, and GCUGCU, and remained when controlling for GpC dinucleotide content. Similar RNA folding analyses of data for RBFOX2 and CELF1 showed a relatively uniform preference for absence of structure at every position across the binding motif, again consistent with predominant binding to single-stranded RNA (Figure 3B).

MBNL Motifs with Unpaired Us Are Associated with Ancient Alternative Exons

In a recent comparative study, we classified conserved exons by their pattern of alternative or constitutive splicing across four mammals and one bird (Merkin et al., 2012) and observed that introns adjacent to exons alternatively spliced in all of the studied mammals (“ancient alternative exons”) are enriched for Mbnl and Rbfox motifs, among others. Curiously, we found that MBNL1 binding to these introns (assayed by CLIP-Seq) exceeded that expected based on motif enrichment by 3- to 4-fold, implying that these introns possess contextual feature(s) that favor binding of MBNL proteins. Performing RNA folding analysis of introns adjacent to exons of different classes, we observed that Us occurring in MBNL motifs such as GCUU that occur near ancient alternative exons have lower base-pairing probability than similar motifs occurring near constitutive exons or more lineage-restricted alternative exons (which showed lower enrichment by CLIP) (Figure 3C). These observations suggest that ancient alternative exons have been selected for presence of MBNL motifs in contexts where the Us are unpaired, likely to facilitate binding by MBNLs.

Motifs Identified In Vitro Are Predominantly Bound In Vivo

To assess the extent to which RBNS motifs are bound in vivo, we used CLIP-Seq data. For RBFOX, a modified version of the high-resolution iCLIP procedure (König et al., 2010) was performed using tagged RBFOX2 in mouse ESCs (mESCs) (Jangi et al., 2014), enabling mapping of sites of crosslinking at nucleotide resolution.

Sites of crosslinking corresponded in many cases to canonical UGCAUG motifs or to the alternate motif, GCACG, identified above. For example, an iCLIP cluster overlapping a GCACG motif was observed in intron 2 of the *Dyrk1a* gene (Figure 4A). To systematically assess the in vivo binding specificity of RBFOX2, the number of crosslinking sites overlapping occurrences of UGCAUG and other motifs in introns and 3′ UTRs were compiled and visualized in a meta-motif representation (Figure 4B). Sharp peaks of crosslinking density directly over UGCAUG sites were present in both introns and 3′ UTRs, illustrating the high specificity of RBFOX2 binding and the high precision of the iCLIP method (Figure 4B, upper). We also observed distinct peaks of crosslink density overlapping occurrences of the alternate motif, GCACG, in both introns and 3′ UTRs (Figure 4B, middle), despite the lack of Us in this motif and the lower abundance of GCACG in the transcriptome (which likely results from presence of a mutation-prone CpG dinucleotide). These peaks were RBFOX2-specific: CLIP-Seq data from an unrelated RBP showed no significant enrichment near canonical or alternate RBFOX2 motifs (Figure 4B, bottom).

Similar analyses of MBNL1 motifs using Mbnl1 CLIP-Seq data from our previously published study with C2C12 mouse myoblasts (Wang et al., 2012) yielded a pronounced peak over Mbnl motifs such as GCUUGC in introns and 3′ UTRs (Figure 4C, upper). Analysis of CELF1 CLIP-Seq data from a study of this factor’s role in splicing and mRNA stability, also using mouse myoblasts (E.T. Wang and C.B.B., unpublished data), yielded a similar peak in the vicinity of canonical CELF motifs such as UGUUGU (Figure 4C, lower). The peaks observed in the MBNL1 and CELF1 CLIP data were not as sharp as those observed for RBFOX2, likely reflecting the lower resolution of the standard CLIP-Seq protocol relative to the iCLIP protocol used for RBFOX2. Again, these peaks were RBP specific (data not shown).

We next compared in vitro and in vivo binding across a broader spectrum of motifs. We defined a CLIP “signal:background” (S/B) ratio for each motif as the total CLIP-Seq read coverage overlapping occurrences of the motif (“signal”) divided by the average of the CLIP coverage in 40 nt regions located at −80 to −41 upstream and +41 to +80 downstream of the motif, representing the background level of CLIP density in motif-containing transcripts. Comparing CLIP S/B values to RBFOX2 RBNS R values across all 6mers, we observed a strong correlation of these values for the set of motifs with significant R values, but not for other 6mers (Figure 4D; left). In fact, 96% of 6mer motifs with significant R value had a CLIP-Seq S/B above the median value for all 6mers (Table S3), including not only all 6mers containing the canonical 5mer GCAUG but also all of those containing the alternate 5mer GCACG. Similar trends were observed for CELF1 and MBNL1, with CLIP-Seq S/B above the median observed for 96% of CELF1 and 99% of MBNL1 6mers with significant R values (Table S3; data for intronic sites in Figure 4D; data for 3′ UTR sites in Figure S5). These observations suggest that the intrinsic binding preferences identified by RBNS determine in vivo binding locations of these proteins to a surprisingly large extent. The observation that virtually all RBNS-enriched

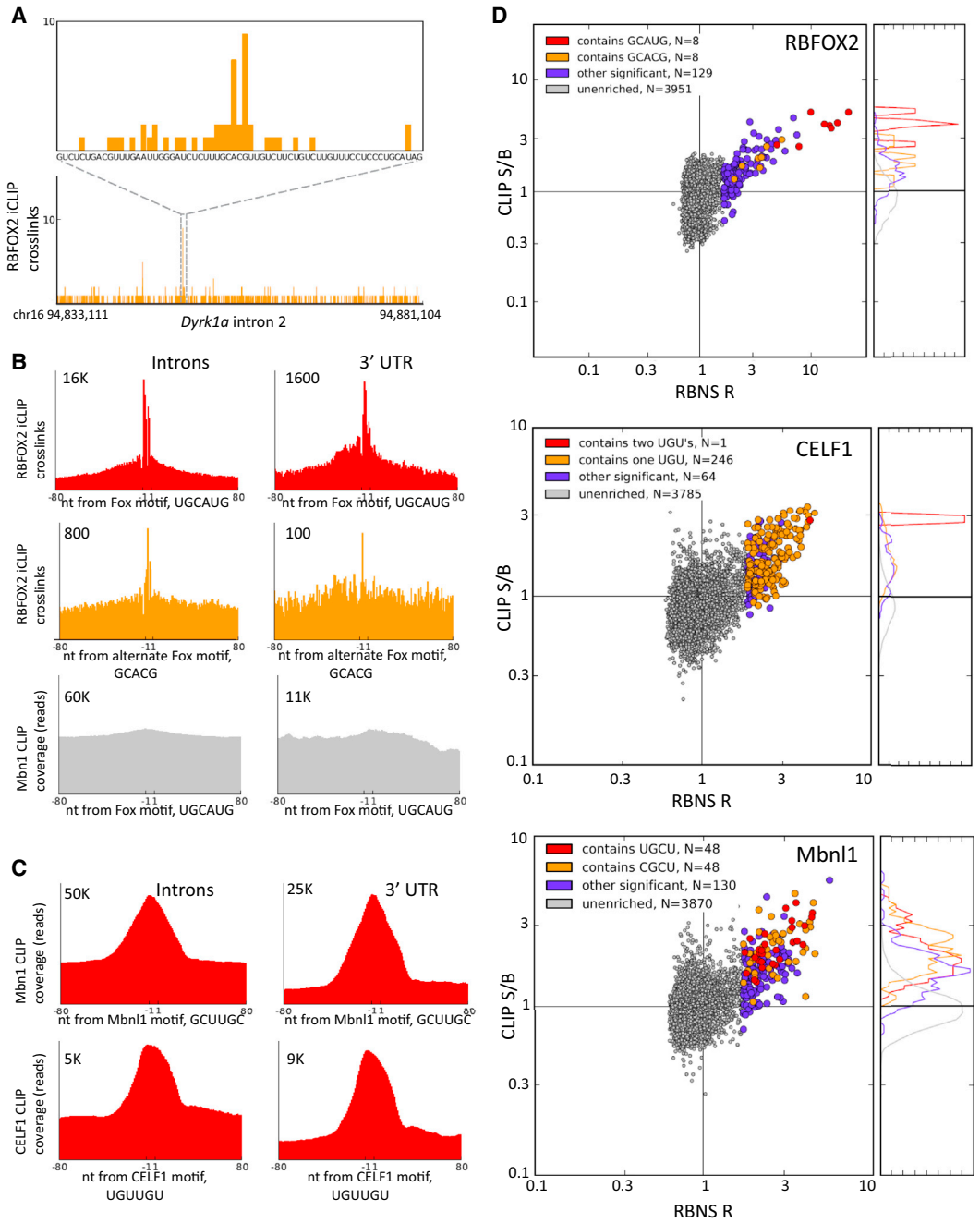


Figure 4. Preferential In Vivo Binding Near RBNS Motifs

(A) The distribution of RBFOX2 iCLIP crosslinking sites (mESCs) in intron 2 of the mouse *Dyrk1a* gene, showing a peak of crosslinks near the alternate motif, GCACG (orange box).

(B) Meta-motif plots (cumulative number of crosslink sites) for RBFOX2 iCLIP data over all occurrences of UGCAUG (top row) in introns (left) and in 3′ UTRs (right), and similarly for the secondary motif GCACG (middle row). The bottom row shows a negative control: meta-motif plot of MBNL1 CLIP data (mouse myoblasts) in the vicinity of the RBFOX motif, UGCAUG. Numbers indicate y axis scale.

(C) Meta-motif plot of MBNL1 CLIP-seq coverage in the vicinity of the top MBNL 6mer, GCUUGC, in introns, and in 3′ UTRs (top row); similar data for CELF1 CLIP-Seq (mouse myoblasts) in the vicinity of the top CELF1 6mer, UUUUGU (bottom row).

(D) Scatter plots of CLIP-Seq S/B versus RBNS R values for each protein analyzed, using same concentrations as in Figure 1, but using 6mers rather than 7mers for CELF1 and MBNL1 to increase statistical power of CLIP S/B analysis. Top: RBFOX2 iCLIP data in introns. Middle: CELF1 CLIP data in introns. Bottom: MBNL1 CLIP data in 3′ UTRs. All significant 6mers containing the indicated submotifs are colored in red, orange, or purple; all nonsignificant 6mers are in gray. Histograms at right show the normalized distributions of CLIP S/B for the corresponding color-coded groups of 6mers.

See also Figure S5.

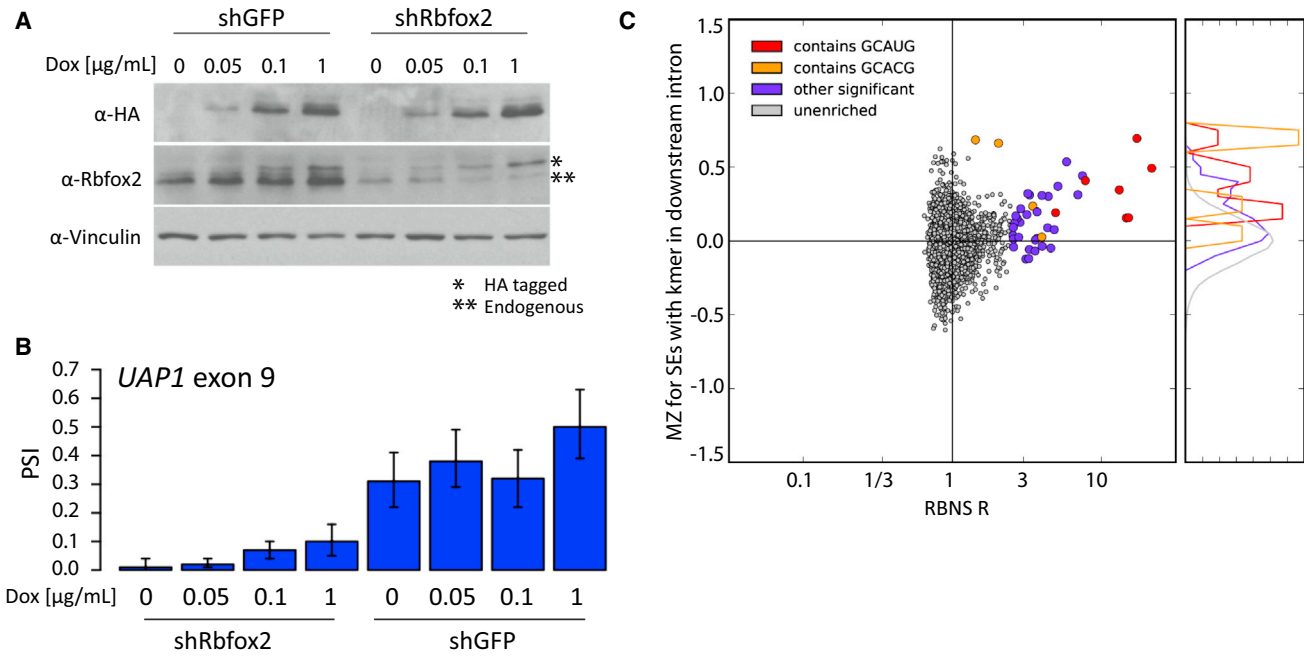


Figure 5. Splicing Regulatory Activity of RNA Motifs from Analysis of Splicing Factor Perturbation Data

(A) Western analysis of Rbfox2 in tet-inducible RBFOX2 mESC lines. Cells were treated with either a control hairpin targeting GFP (left lanes) or a hairpin targeting endogenous *Rbfox2* mRNAs (right lanes). Cells were treated with 0, 0.05, 0.1, or 1 μg/mL of Dox to induce exogenous FLAG-tagged RBFOX2. Western shows endogenous and tagged Rbfox2 as well as a loading control (Vinculin).

(B) The percent spliced in (PSI) values shown for a highly Rbfox2-sensitive alternative exon in pyrophosphorylase *Uap1* in mESCs at each of the eight different Rbfox2 levels shown above (two hairpins × four levels of Dox). Error bars show 95% confidence intervals.

(C) Distribution of RBFOX2 monotonicity Z scores versus RBFOX2 RBNS R values for all 6mers. MZ scores were calculated for 1,442 skipped exons in mESC-expressed genes using the Rbfox2 perturbation system shown in (A). For each 6mer, the average MZ score of all exons that had the 6mer in the first 200 bases of the downstream intron was calculated. Coloring as in Figure 4. RBNS-enriched 6mers had significantly higher MZ scores than unenriched 6mers (KS test, $p = 2 \times 10^{-7}$).

See also Figure S6.

motifs had CLIP signal above the median suggests that a substantial majority of motifs detected in vitro by RBNS are bound in vivo to at least some extent. However, this relationship was not reciprocal: many motifs with high CLIP S/B were bound in vitro, but many others lacked significant in vitro binding, a phenomenon that we explore below.

Alternate and Canonical Motifs Are Associated with Alternative Splicing Regulation

To explore the splicing regulatory activity of the RBFOX2 motifs identified by RBNS, mESCs with a range of RBFOX2 expression levels were generated. Overexpression of RBFOX2 to different extents was achieved by administration of various concentrations of doxycycline to a mESC line containing a tetracycline-inducible version of RBFOX2 (Jangi et al., 2014). Inhibition of RBFOX2 expression was achieved by stably introducing vectors expressing short hairpin RNAs (shRNAs) targeting the 3′ UTR of the endogenous gene (or shRNAs targeting GFP as a control). RNA-Seq analysis of cell lines expressing eight different levels of RBFOX2 proteins was then performed to assess changes in alternative splicing.

Expression of *Rbfox2* increased from 12 fragments per kilobase of exon per million mapped fragments (FPKM) in the lowest condition (shFOX2, 0 μg/mL DOX) to an FPKM of 32 at the highest induced level (shGFP, 1 μg/mL Dox), ranging from 40% to 123% of endogenous levels, which is still lower than occurs in certain mouse tissues (Figure S6). Protein levels were confirmed by western analysis (Figure 5A). To systematically assess the consistency of changes in splicing, we defined a "monotonicity Z score" (MZ) for each exon whose "percent spliced in" (PSI) value changed significantly (E.T. Wang and C.B.B., unpublished data). MZ captures the extent to which the exon's PSI consistently increases (MZ > 0) or consistently decreases (MZ < 0) in a set of conditions with increasing levels of a regulatory factor, as is expected to occur for direct regulatory targets.

Applying this approach to a set of mouse alternative exons, the exons with the highest MZ scores were exon 9 of the *UAP1* gene (MZ = 2.98) and the EIIIB exon of Fibronectin1 (MZ = 2.81). The latter is a well-established Rbfox2 target whose downstream intron contains six canonical UGCAUG motifs (Huh and Hynes, 1993; Jin et al., 2003; Lim and Sharp, 1998). RNA-Seq data for the regulated *UAP1* exon are displayed in Figure 5B, showing that the PSI value increases from below 10% in conditions where Rbfox2 is depleted to 61% in the highest overexpression condition. To assess the extent to which particular

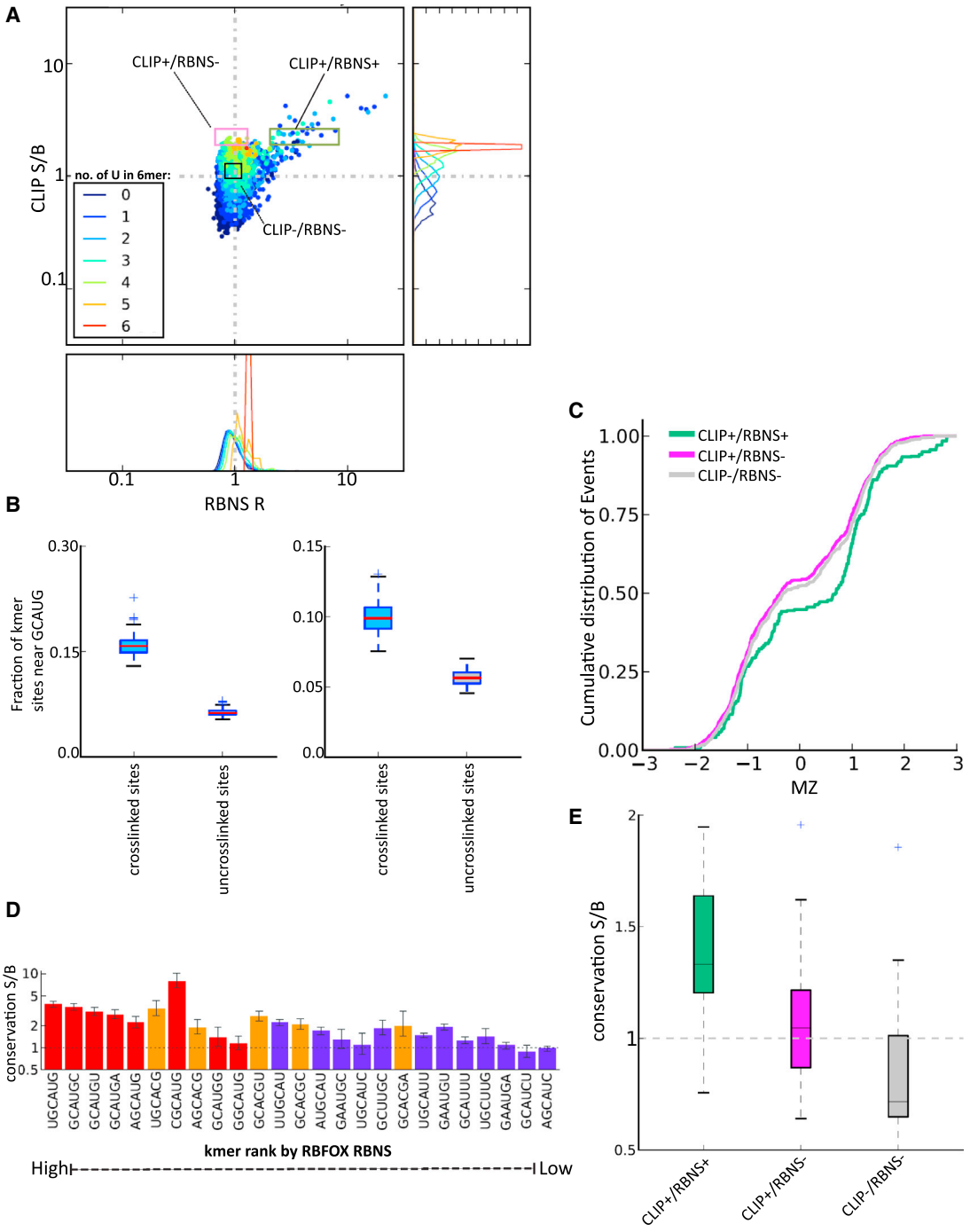


Figure 6. RBNS Distinguishes Subsets of CLIP-seq Motifs with and without Regulatory Activity

(A) RBFOX2 iCLIP S/B in 3′ UTRs is plotted against RBFOX2 RBNS R value for all 6mers (as in Figure 4D), with points colored by the number of U bases present in the 6mer as indicated. The distribution of iCLIP S/B values is shown at right, and the distribution of RBNS R values are shown below, for each group of 6mers binned by U content. Log scale is used on both axes.

(B) RBFOX primary motifs have increased frequency near crosslinked CLIP+/RBNS− sequences. For each CLIP+/RBNS− motif in either introns (left) or 3′ UTRs (right), the fraction of motifs that had a GCAUG within 40 nt was calculated for all motif occurrences that were crosslinked in iCLIP or uncrosslinked. Boxplots show the fraction of CLIP+/RBNS- sites that are near a canonical motif.

(C) Cumulative distribution of MZ scores for sets of alternative exons grouped by presence of specific 6mer motifs in first 200 nt of downstream intron. Groups of 6mers are colored as in (A).

(legend continued on next page)

sequence motifs were associated with splicing regulation, we defined an MZ score for each 6mer as the average MZ value of alternative exons that have the 6mer present in the first 200 bases of the downstream intron, a region in which RBFOX2 binding is associated with activation of exon inclusion (Ponthier et al., 2006; Yeo et al., 2009). Comparing motif MZ scores with RBNS R values of 6mers, we observed that >80% of 6mers with significant R values had positive MZ scores, consistent with a role in enhancement of splicing in response to increased RBFOX2 levels (Figure 5C). Positive MZ scores were observed not only for all 6mers containing the canonical GCAUG 5mer but also for all 6mers containing the GCACG alternate motif, supporting that this motif confers RBFOX-dependent splicing regulation.

RBNS Detects Sequence Bias in CLIP Data

CLIP-Seq is a widely used and effective technique for mapping RBP binding sites in vivo (Sugimoto et al., 2012). However, the absence of alternative comprehensive high-resolution methods for measuring in vivo binding has made it difficult to critically assess CLIP data for systematic biases or sources of false positives and false negatives. Previous studies have shown that CLIP favors U-rich sequences, because uridines form RNA-protein crosslinks more readily than other bases (Sugimoto et al., 2012). Coloring 6mers according to the number of Us that they contained in the plot of RBFOX2 CLIP S/B against RBNS R values revealed a group of 6mers with high U content ($\geq$4 U out of 6) at the top center of the distribution with high CLIP S/B but no significant RBNS enrichment (Figure 6A). By contrast, the remainder of 6mers with high CLIP S/B also had significant positive RBNS R values and contained moderate numbers of Us (usually one or two). This observation and the systematic trend for higher iCLIP S/B values to be associated with higher U content (Figure 6A; right) suggested that U richness systematically and substantially enhances detection by CLIP, to an extent that essentially nonspecific (low specificity) protein-RNA interactions may be detected in contexts that are sufficiently U rich.

To determine the extent to which CLIP+/RBNS− motifs result from binding to U-rich sequences near authentic RBFOX motifs, we analyzed the sequences surrounding crosslinked CLIP+/RBNS− motifs (Figure 6B). While we observed a ~2-fold increase in GCAUG occurences near these sites (within 40 nt) relative to uncrosslinked occurrences of these motifs, presence of a nearby GCAUG motif was observed for only ~15% of crosslinked sites associated with CLIP+/RBNS− motifs (Figure 6B). These data suggest that some CLIP signal for such motifs comes from binding to nearby canonical motifs, but most such binding derives from crosslinking of protein that is associated with RNA nonspecifically or via interaction with other RBPs.

To assess the splicing activity of motifs detected exclusively by CLIP, we compared the splicing regulatory activity of three sets of motifs: (i) 6mers with high CLIP S/B but low RBNS R values (the CLIP+/RBNS− set), (ii) 6mers with significant RBNS R values and CLIP S/B values in the same range as the previous set (CLIP+/RBNS+), and (iii) a negative control group of sequences that lacked enrichment by CLIP or RBNS (CLIP−/RBNS−) (Figure 6A). Comparing the splicing regulation of cassette exons whose downstream introns contain 6mers from each set revealed a clear pattern: exons associated with the CLIP+/RBNS+ set had significantly higher MZ scores than those associated with either control 6mers or with CLIP+/RBNS− 6mers. Furthermore, the CLIP+/RBNS− set was no more likely to be associated with high MZ values than the control set (Figure 6C). Thus, no evidence was found that the CLIP+/RBNS− set of motifs has regulatory activity. Instead, the simplest explanation is that these motifs result from transient nonspecific interactions of protein with RNA, with U-rich sequences preferentially captured relative to other nonspecifically bound RNAs. This analysis shows that RBNS can provide information useful for interpretation of CLIP-Seq data. On the other hand, the observation that essentially all significant RBNS 6mers also had high CLIP S/B values argues against the existence of a class of CLIP-invisible (e.g., uncrosslinkable) RNA motifs, at least for RBFOX2.

RBNS Motifs Are Conserved Across Mammals

Motifs that contribute to regulation of conserved alternative splicing events should often be evolutionarily conserved, and the canonical binding motifs of RBFOX2, MBNL1, and CELF1 are highly conserved in introns flanking alternative exons and in 3′ UTRs (Daughters et al., 2009; Merkin et al., 2012; Sugnet et al., 2006; Wang et al., 2008, 2012). Adapting a method previously developed to assess conservation of microRNA target sites in mRNAs (Friedman et al., 2009), we assessed the conservation of significant RBFOX2 RBNS motifs in orthologous UTRs of 23 mammalian species. UTRs were chosen over introns because they can be more reliably aligned in most cases. For this analysis, we calculated for each 6mer the fraction of its occurrences in conserved introns that were evolutionarily conserved over at least a minimum evolutionary branch length (the "signal"). We measured a similar fraction for a cohort of control 6mers matched for genomic abundance, C+G% and CpG dinucleotide content, defining the mean conserved fraction over these control 6mers as the "background." For RBFOX motifs, almost all 6mers containing the canonical GCAUG 5mer had conservation S:B ratio significantly above 1, indicating preferential conservation (Figure 6D). Furthermore, 6mers containing the alternative motif GCACG had S:B values nearly as high, further supporting the in vivo regulatory function of this motif. Some but not all of the remaining RBNS motifs also had significant S:B values, supporting function. No significant conservation was detected for the set of CLIP+/RBNS− 6mers (Figure 6E), consistent with lack of regulatory activity. By contrast, the set of CLIP+/RBNS+ motifs matched for CLIP density showed significant conservation (Figure 6E).

(D) Conservation S/B of the top RBFOX2 6mer motifs by RBNS in mammalian 3′ UTRs. Motifs are listed in descending order of R value and colored as in previous figures. Error bars indicate 95% confidence intervals generated by resampling background *k*mers.
(E) Box plots of the distributions of conservation S/B for 6mers grouped as in (A): CLIP+/RBNS+, CLIP+/RBNS−, and CLIP−/RBNS−. Conservation S/B was calculated as in (D).

DISCUSSION

The RBNS method and associated analytical approaches that provide comprehensive and quantitative information about the spectrum of RNA motifs bound by an RBP. As affinities for all *k*mers are assessed simultaneously, this approach may prove attractive as an alternative to traditional low-throughput quantitative methods. To address more targeted questions related to specific RBPs, various details of the RBNS experimental setup could be varied, including the length or composition of the input RNA or the presence of additional protein factors that are hypothesized to cooperate or compete with the protein being pulled down. Instead of random RNA, total cellular RNA, mRNA, or RNA immunoprecipitated with an RBP could be used to limit sampled sequences to potential in vivo binding sites. This approach could enable detection of binding to sites with complex architecture engineered by evolution but would substantially reduce sequence diversity, limiting the power to analyze binding to longer motifs or effects of RNA structure. Current sequencing technologies limit motif size to about ten bases, but there are strategies to circumvent this limit (Supplemental Experimental Procedures).

Complexity of RNA Binding Affinity Spectra

The depth of data generated in this approach yields information across a broad range of binding affinities, particularly when several RBP concentrations are used, enabling detection of weaker but significant motifs, such as GCACG for RBFOX2. For this particular example, the structure of the RBFOX1 RRM domain (which is identical to that of RBFOX2) has been solved by NMR, in complex with RNA representing the canonical motif, UGCAUG (Auweter et al., 2006). The substitution of U for C in the fifth position of the 6mer would not introduce a steric clash, and one of the two hydrogen bonds that RBFOX1 makes with U5 would be preserved with a C in this position (Auweter et al., 2006). Together, these observations suggest that RBFOX proteins can bind GCACG in a manner similar to their binding of GCAUG, albeit with somewhat lower affinity. These observations, and similar results for a variety of variants of classical CELF1 and MBNL1 motifs, lead us to conclude that RBPs often have rather complex RNA binding affinity spectra, often centered on core dinucleotides, such as the GUs and GCs present in CELF1 and MBNL1 motifs, respectively. We also found that GCACG motifs are bound in vivo and are associated with sequence conservation and splicing regulatory activity to an extent similar to canonical motifs. These and similar observations for a variety of variant CELF1 and MBNL1 motifs argue that secondary motifs with affinities within an order of magnitude or so of the optimal motif often play conserved roles in splicing regulation.

We envision several types of applications for RBNS and the resulting data. These applications include modeling and predicting changes in RBP occupancy and regulatory activity in response to changes in RBP abundance or activity occurring during development, between cell types, or in different cell states (e.g., EMT and disease versus normal) and predicting the regulatory consequences of genetic variation (e.g., disease gene mutations or polymorphisms) on RBP binding and regulatory activity. For these applications, the quantitative precision of the F_i and K_d values from the SKA algorithm may prove useful. Other potential applications include understanding the influence of RNA secondary structure on RBP binding and function and interpreting CLIP-Seq data. These last two applications are discussed below.

Effects of Structure on RNA Binding

The impact of RNA structure on protein-RNA interactions can be inferred using RBNS. For RBFOX2 and CELF1, both of which bind RNA through RRM domains, our RNA folding analyses suggested strong preferences for binding of single-stranded RNA. Analysis of MBNL1, which binds RNA through zinc fingers, revealed a strong preference for unpaired Us but no significant bias for or against unpaired G and C bases in UGC-containing motifs, suggesting either that MBNL can melt paired GC dinucleotides or that it can recognize them even when they are base-paired. CUG repeat RNA, which is tightly bound by MBNL proteins both in vitro and in vivo (Teplova and Patel, 2008), crystallizes as a hairpin with paired GCs separated by unpaired U-U bulges (Mooers et al., 2005), consistent with the pattern of MBNL binding preferences observed here. Intron 4 of cardiac troponin T (*cTNT*), a well-characterized MBNL binding and regulatory target, also contains multiple paired GCs flanked by unpaired pyrimidine bulges (Warf and Berglund, 2007). Consistently, biochemical evidence has shown that MBNL binds with high affinity to pairs of GC dinucleotides with a wide range (~1–15) of intervening pyrimidine bases (Goers et al., 2010; Cass et al., 2011). This structural signature is consistent with RNA looping around MBNL proteins such that different zinc fingers interact with different GCs. RNA looping as a mechanism of RNA recognition has been proposed for PTB (Oberstrass et al., 2005; Pérez et al., 1997) and is also consistent with the crystal structure of MBNL1 zinc fingers 3 and 4 (Teplova and Patel, 2008).

RBNS Enhances Interpretation of CLIP Data

RBNS appears to yield a less biased portrait of the spectrum of RNA motifs bound by an RBP than do methods based on UV crosslinking, making it a useful complement to CLIP-based methods (including iCLIP and PAR-CLIP). The subset of CLIP-enriched motifs that were not detected by RBNS lacked evidence of regulatory activity or sequence conservation, arguing that they do not reflect biologically relevant binding. In practice, when crosslinking to a CLIP+/RBNS− motif that is located in close proximity to a CLIP+/RBNS+ motif is observed, our analyses imply that in most cases this binding should be attributed to the CLIP+/RBNS+ motif. Applying this sort of correction automatically might improve inference of regulatory elements. When comparing the extent of binding to two or more different regions, we expect that RBNS affinities could be used to correct for the crosslinking bias inherent in CLIP and improve the accuracy of quantitation.

EXPERIMENTAL PROCEDURES

Cloning, Expression, and Purification of Proteins

Full-length *CELF1*, *MBNL1* (1–260), and *RBFOX2* (100–194) were cloned downstream of a GST-SBP tandem affinity tag. Both truncated *MBNL1* and

RBFOX2 constructs contain all RNA binding domains, including all four *MBNL1* zinc finger domains and *RBFOX2*'s single RRM. The proteins were recombinantly expressed, purified via the GST tag, and the GST tag cleaved off with Prescission protease (GE).

Bioanalyzer Analysis

For each protein concentration in the RBFOX RBNS experiment, the amount of RBP-bound RNA was measured using a Bioanalyzer (Agilent Technologies). RNA extracted from each RBNS concentration was run on a RNA 6000 pico chip for low- and no-RBP conditions or a RNA 6000 nano chip for high-RBP concentrations (Agilent Technologies). These measurements were made according to the manufacturer's instructions and were used to estimate the concentration of RBFOX2 in complex with RNA in order to calculate relative RBNS binding affinities (see Supplemental Information).

RBNS

RBNS was performed after purifying a given RBP and in vitro transcribing RBNS input RNA; experimental details can be found in Supplemental Experimental Procedures. Seven to ten concentrations of RBP, including a no RBP condition, was equilibrated in binding buffer for 30 min at room temperature or 37°C in the case of RBFOX RBNS. RBNS input random RNA was then added to a final concentration of 1 μM with 40 U of Superasin (Ambion) and incubated for 1 hr at room temperature or 37°C. To pull down tagged RBP and interacting RNA, each RNA/protein solution was then added to 1 mg of washed streptavidin magnetic beads and incubated for 1 hr. Unbound RNA was removed from the beads, and the beads were washed once with 1 ml of wash buffer. The beads were incubated at 70°C for 10 min in 100 μl of elution buffer (10 mM tris [pH 7.0], 1mM EDTA, and 1% SDS), and the eluted material was collected. Bound RNA was extracted, reverse transcribed into cDNA, and then amplified by PCR. See Supplemental Experimental Procedures for a more detailed description of the RBNS protocol.

R Values

Motif R values were calculated as the motif frequency in the RBP-selected pool over the frequency in the input RNA library. Frequencies were controlled for respective library read depth. R values were considered significant if greater than 2 SDs from the mean. The rate of kmer enrichment in the no protein condition, relative to the input library, was defined as the FDR.

SKA Analysis

The streaming *k*mer assignment algorithm is described in Supplemental Experimental Procedures. See also Figures S2 and S3.

Monotonicity Z Scores

Each of the eight RNA-seq libraries was mapped to the mouse genome (mm9) with Tophat, and the alternative splicing of skipped exon (SE) events was analyzed with MISO (Katz et al., 2010) as follows: significantly changing (Bayes factor $\geq$5.0) events were identified from all pairwise comparisons between the libraries. The difference between the number of comparisons where the higher RBFOX concentration showed significantly more inclusion and the number where the lower RBFOX concentration showed more inclusion was calculated for all events. For each skipped exon event, the monotonicity score was defined to be the Z score of this difference out of a control set of differences generated by shuffling the order of the RBFOX concentration data sets.

Software

All software described here is freely available for academic use on github (https://github.com/alexrson/rbns_pipeline).

ACCESSION NUMBERS

RBNS and RNA-seq sequencing data have been deposited into the Sequence Read Archive (SRA) under the accession number SRP041098.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, three tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.04.016.

ACKNOWLEDGMENTS

We thank Andy Berglund for advice on expression of RBPs; Eric Wang for helpful discussions; and Tom Cooper, Wendy Gilbert, and Andy Berglund for helpful suggestions on the text. We also thank Vincent Butty for his help with conservation analyses, Avery Whitlock for artwork (Figure 1A), and Albert Tai for performing SPR analysis. This work was funded by an NIH NRSA Postdoctoral Fellowship (N.L.), by grant number 0821391 from the National Science Foundation, and by grants from the NIH (C.B.B.).

Received: December 27, 2013
Revised: March 4, 2014
Accepted: April 10, 2014
Published: May 15, 2014

REFERENCES

Auweter, S.D., Fasan, R., Reymond, L., Underwood, J.G., Black, D.L., Pitsch, S., and Allain, F.H. (2006). Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. EMBO J. *25*, 163–173.

Baraniak, A.P., Chen, J.R., and Garcia-Blanco, M.A. (2006). Fox-2 mediates epithelial cell-specific fibroblast growth factor receptor 2 exon choice. Mol. Cell. Biol. *26*, 1209–1222.

Campbell, Z.T., Bhimsaria, D., Valley, C.T., Rodriguez-Martinez, J.A., Menichelli, E., Williamson, J.R., Ansari, A.Z., and Wickens, M. (2012). Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. Cell Rep. *1*, 570–581.

Cass, D., Hotchko, R., Barber, P., Jones, K., Gates, D.P., and Berglund, J.A. (2011). The four Zn fingers of MBNL1 provide a flexible platform for recognition of its RNA binding elements. BMC Mol. Biol. *12*, 20.

Daughters, R.S., Tuttle, D.L., Gao, W., Ikeda, Y., Moseley, M.L., Ebner, T.J., Swanson, M.S., and Ranum, L.P. (2009). RNA gain-of-function in spinocerebellar ataxia type 8. PLoS Genet. *5*, e1000600.

Edwards, J.M., Long, J., de Moor, C.H., Emsley, J., and Searle, M.S. (2013). Structural insights into the targeting of mRNA GU-rich elements by the three RRMs of CELF1. Nucleic Acids Res. *41*, 7153–7166.

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

Gehman, L.T., Meera, P., Stoilov, P., Shiue, L., O'Brien, J.E., Meisler, M.H., Ares, M., Jr., Otis, T.S., and Black, D.L. (2012). The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. Genes Dev. *26*, 445–460.

Goers, E.S., Purcell, J., Voelker, R.B., Gates, D.P., and Berglund, J.A. (2010). MBNL1 binds GC motifs embedded in pyrimidines to regulate alternative splicing. Nucleic Acids Res. *38*, 2467–2484.

Ho, T.H., Charlet-B, N., Poulos, M.G., Singh, G., Swanson, M.S., and Cooper, T.A. (2004). Muscleblind proteins regulate alternative splicing. EMBO J. *23*, 3103–3112.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. *31*, 3429–3431.

Huh, G.S., and Hynes, R.O. (1993). Elements regulating an alternatively spliced exon of the rat fibronectin gene. Mol. Cell. Biol. *13*, 5301–5314.

Jangi, M., Boutz, P.L., Paul, P., and Sharp, P.A. (2014). Rbfox2 controls autoregulation in RNA-binding protein networks. Genes Dev. *28*, 637–651.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. EMBO J. *22*, 905–912.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Kalsotra, A., Xiao, X., Ward, A.J., Castle, J.C., Johnson, J.M., Burge, C.B., and Cooper, T.A. (2008). A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. Proc. Natl. Acad. Sci. USA *105*, 20333–20338.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.

Kuyumcu-Martinez, N.M., Wang, G.S., and Cooper, T.A. (2007). Increased steady-state levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. Mol. Cell *28*, 68–78.

Ladd, A.N., Charlet, N., and Cooper, T.A. (2001). The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol. Cell. Biol. *21*, 1285–1296.

Lim, L.P., and Sharp, P.A. (1998). Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats. Mol. Cell. Biol. *18*, 3900–3906.

Mankodi, A., Lin, X., Blaxall, B.C., Swanson, M.S., and Thornton, C.A. (2005). Nuclear RNA foci in the heart in myotonic dystrophy. Circ. Res. *97*, 1152–1155.

Marquis, J., Paillard, L., Audic, Y., Cosson, B., Danos, O., Le Bec, C., and Osborne, H.B. (2006). CUG-BP1/CELF1 requires UGU-rich sequences for high-affinity binding. Biochem. J. *400*, 291–301.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593–1599.

Mooers, B.H., Logue, J.S., and Berglund, J.A. (2005). The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. Proc. Natl. Acad. Sci. USA *102*, 16626–16631.

Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol. *29*, 659–664.

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L., and Allain, F.H. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science *309*, 2054–2057.

Pérez, I., McAfee, J.G., and Patton, J.G. (1997). Multiple RRMs contribute to RNA binding specificity and affinity for polypyrimidine tract binding protein. Biochemistry *36*, 11881–11890.

Ponthier, J.L., Schluepen, C., Chen, W., Lersch, R.A., Gee, S.L., Hou, V.C., Lo, A.J., Short, S.A., Chasis, J.A., Winkelmann, J.C., and Conboy, J.G. (2006). Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. J. Biol. Chem. *281*, 12468–12474.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. *27*, 667–670.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods *10*, 71–73.

Stefl, R., Skrisovska, L., and Allain, F.H. (2005). RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. *6*, 33–38.

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

Sugnet, C.W., Srinivasan, K., Clark, T.A., O'Brien, G., Cline, M.S., Wang, H., Williams, A., Kulp, D., Blume, J.E., Haussler, D., and Ares, M., Jr. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput. Biol. *2*, e4.

Taneja, K.L., McCurrach, M., Schalling, M., Housman, D., and Singer, R.H. (1995). Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. J. Cell Biol. *128*, 995–1002.

Teplova, M., and Patel, D.J. (2008). Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. Nat. Struct. Mol. Biol. *15*, 1343–1351.

Timchenko, L.T., Miller, J.W., Timchenko, N.A., DeVore, D.R., Datar, K.V., Lin, L., Roberts, R., Caskey, C.T., and Swanson, M.S. (1996). Identification of a (CUG)n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. Nucleic Acids Res. *24*, 4407–4414.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

Underwood, J.G., Boutz, P.L., Dougherty, J.D., Stoilov, P., and Black, D.L. (2005). Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. Mol. Cell. Biol. *25*, 10005–10016.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, E.T., Cody, N.A., Jog, S., Biancolella, M., Wang, T.T., Treacy, D.J., Luo, S., Schroth, G.P., Housman, D.E., Reddy, S., et al. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell *150*, 710–724.

Warf, M.B., and Berglund, J.A. (2007). MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T. RNA *13*, 2238–2251.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat. Struct. Mol. Biol. *16*, 130–137.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Res. *37*, e151.

Appendix C.

mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues

Stephen W. Eichhorn[1,2,3]*, Huili Guo[1,2,3,4,5,6]*, Sean E. McGeary[1,2,3], Ricard A. Rodriguez-Mias[7], Chanseok Shin[1,2,8], Daehyun Baek[1,2,9,10,11], Shu-hao Hsu[12], Kalkpana Ghoshal[12], Judit Villén[7], and David P. Bartel[1,2,3]

[1]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Institute of Molecular and Cell Biology, Singapore 138673, Singapore
[5]Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore
[6]Lee Kong Chain School of Medicine, Nanyang Technological University-Imperial College, Singapore 639798, Singapore
[7]Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA
[8]Department of Agricultural Biotechnology, Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-921, Republic of Korea
[9]Center for RNA Research, Institute for Basic Science, Seoul 151-747, Republic of Korea
[10]School of Biological Sciences, Seoul National University, Seoul 151-747, Republic of Korea
[11]Bioinformatics Institute, Seoul National University, Seoul 151-742, Republic of Korea
[12]Department of Pathology, Ohio State University, Columbus, OH 43210, USA
*These authors contributed equally to this work.

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

CellPress

Molecular Cell
Article

mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues

Stephen W. Eichhorn,[1,2,3,13] Huili Guo,[1,2,3,4,5,6,13] Sean E. McGeary,[1,2,3] Ricard A. Rodriguez-Mias,[7] Chanseok Shin,[1,2,8] Daehyun Baek,[1,2,9,10,11] Shu-hao Hsu,[12] Kalpana Ghoshal,[12] Judit Villén,[7] and David P. Bartel[1,2,3,*]

[1]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Institute of Molecular and Cell Biology, Singapore 138673, Singapore
[5]Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore
[6]Lee Kong Chian School of Medicine, Nanyang Technological University-Imperial College, Singapore 639798, Singapore
[7]Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA
[8]Department of Agricultural Biotechnology, Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-921, Republic of Korea
[9]Center for RNA Research, Institute for Basic Science, Seoul 151-747, Republic of Korea
[10]School of Biological Sciences, Seoul National University, Seoul 151-747, Republic of Korea
[11]Bioinformatics Institute, Seoul National University, Seoul 151-742, Republic of Korea
[12]Department of Pathology, Ohio State University, Columbus, OH 43210, USA
[13]Co-first author
*Correspondence: dbartel@wi.mit.edu
http://dx.doi.org/10.1016/j.molcel.2014.08.028

SUMMARY

MicroRNAs (miRNAs) regulate target mRNAs through a combination of translational repression and mRNA destabilization, with mRNA destabilization dominating at steady state in the few contexts examined globally. Here, we extend the global steady-state measurements to additional mammalian contexts and find that regardless of the miRNA, cell type, growth condition, or translational state, mRNA destabilization explains most (66%–>90%) miRNA-mediated repression. We also determine the relative dynamics of translational repression and mRNA destabilization for endogenous mRNAs as a miRNA is induced. Although translational repression occurs rapidly, its effect is relatively weak, such that by the time consequential repression ensues, the effect of mRNA destabilization dominates. These results imply that consequential miRNA-mediated repression is largely irreversible and provide other insights into the nature of miRNA-mediated regulation. They also simplify future studies, dramatically extending the known contexts and time points for which monitoring mRNA changes captures most of the direct miRNA effects.

INTRODUCTION

MicroRNAs (miRNAs) are small, noncoding RNAs that posttranscriptionally regulate the expression of most mammalian genes (Bartel, 2009; Friedman et al., 2009). Acting as the specificity components of ribonucleoprotein silencing complexes, miRNAs pair with target mRNAs at sites complementary to the miRNA 5′ region. Most effective sites map to 3′ untranslated regions (3′ UTRs) and pair perfectly with the miRNA seed (nucleotides 2–7), with an additional pair at nucleotide 8 and/or an A across from nucleotide 1 (Bartel, 2009).

Although early reports of gene regulation by miRNAs emphasized their role as translational repressors (Wightman et al., 1993; Olsen and Ambros, 1999; Seggerson et al., 2002), subsequent studies revealed that miRNAs can also induce mRNA degradation (Bagga et al., 2005; Krützfeldt et al., 2005; Lim et al., 2005). This degradation is a consequence of miRNA-mediated deadenylation of target mRNAs (Behm-Ansmant et al., 2006; Giraldez et al., 2006; Wu et al., 2006), which causes these mRNAs to undergo decapping and then 5′–3′ decay (Rehwinkel et al., 2005; Behm-Ansmant et al., 2006; Chen et al., 2009). The discovery of this second mode of repression raised the question as to the relative contributions of translational repression and mRNA degradation to reducing the protein abundance of regulated genes.

Large-scale analyses comparing protein and mRNA changes of predicted miRNA targets after introducing or deleting individual mammalian miRNAs found that protein changes generally correspond to changes in polyadenylated mRNA abundance (Baek et al., 2008). More precise measurements comparing changes in translational efficiency (TE) to changes in mRNA again found that mRNA degradation explains the majority of miRNA-mediated repression, with translational repression contributing roughly 10%–25% of the overall repression (Hendrickson et al., 2009; Guo et al., 2010). These global measurements of TE and mRNA (or protein and mRNA) were made at relatively late time points (12–32 hr after introducing the miRNA

or long after induction of an endogenous miRNA) and thus are thought to reflect the steady-state effects of the miRNA (Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010). When miRNAs are expressed at constant levels, steady-state measurements are ideal for quantifying the relative contributions of translational repression and mRNA degradation because they integrate effects occurring throughout the life cycle of each targeted transcript.

If generalizable to other cell types and conditions, these high-throughput steady-state measurements, which indicate that mRNA changes closely approximate the overall effects of a miRNA on target gene expression, would be welcome news for those placing mammalian miRNAs into gene regulatory networks and quantifying their impact on gene expression, since measuring changes in mRNA levels is much easier than measuring changes in protein levels or TE. However, protein/TE and mRNA effects have been globally compared in only two cell lines, HeLa cells (Baek et al., 2008; Selbach et al., 2008; Guo et al., 2010) and human embryonic kidney 293T (HEK293T) cells (Hendrickson et al., 2009), and a single primary cell type, mouse neutrophils (Baek et al., 2008; Guo et al., 2010), which leaves open the possibility that translational repression might dominate in most other mammalian contexts.

The observation that mRNA destabilization can account for most repression at steady state has prompted a search for time points in which translational repression might explain a larger proportion of the repression. Two studies examined the dynamics of miRNA-mediated repression on inducible reporter genes as these genes begin to be expressed in fly and human cells (Béthune et al., 2012; Djuranovic et al., 2012), and another examined the effects of miR-430 on its endogenous targets in the zebrafish embryo (Bazzini et al., 2012). In blastula-stage zebrafish embryos (4 hr postfertilization [hpf]), miR-430 substantially reduces the TE of its targets with little effect on their stability, whereas by gastrulation (6 hpf), the relative contributions of TE and mRNA destabilization closely resemble those observed previously at steady state in mammalian systems (Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010; Bazzini et al., 2012). Because miR-430 is strongly induced shortly before the blastula stage, the large amount of translational repression observed in the blastula stage, followed by the mRNA destabilization observed later in the gastrula stage, was proposed to reflect the fundamental dynamics of miRNA-mediated repression (Bazzini et al., 2012).

The idea that miRNA-mediated translational repression precedes mRNA degradation cannot be disputed—an mRNA molecule can undergo translational repression only before it has been degraded, and thus its translational regulation must precede regulation at the level of its stability in the same way that transcriptional regulation must precede translational regulation. However, subsequent insight into the shift in regulatory regime occurring as zebrafish embryos progress from pre- to postgastrulation has overturned the idea that the miR-430 observations reflect the dynamics of miRNA-mediated repression (Subtelny et al., 2014). Prior to gastrulation, mRNA poly(A) tail length and TE are coupled, and short-tailed mRNAs are stable. These two unique conditions enable miRNA-mediated deadenylation to cause translational repression without mRNA destabilization (Subtelny et al., 2014). The transition to mostly mRNA decay is due to a change in these conditions at gastrulation such that coupling between tail length and TE is lost and short-tailed mRNAs become less stable, which causes the consequence of miRNA-mediated deadenylation to shift from translational repression to mRNA destabilization (Subtelny et al., 2014). When considering this shifting regulatory regime, the miR-430 results do not provide insight into the dynamics of the two modes of miRNA-mediated repression for endogenous mRNAs, nor do they demonstrate that miRNA-mediated translational repression occurring through a deadenylation-independent mechanism ever mediates meaningful changes in the expression of endogenous mRNAs. This being said, the miR-430 study is notable in that it identified an endogenous setting in which the effects of a miRNA cannot be approximated by changes in mRNA levels (Bazzini et al., 2012). Because of the regulatory regime operating in the pregastrulation zebrafish embryo (and presumably in other early embryos or other unusual settings, such as neuronal synapses), measuring mRNA changes misses essentially all of the effects of miRNAs in this setting (Subtelny et al., 2014).

The two studies that monitor reporter genes rather than endogenous transcripts to examine miRNA repression dynamics both report that a phase of substantial translational repression occurs prior to detectable mRNA deadenylation or decay (Béthune et al., 2012; Djuranovic et al., 2012). However, the updated understanding of the miR-430 results reopens the question of whether such a phase also occurs for endogenous mRNAs. Although reporters can faithfully represent endogenous genes, several observations led us to suspect that when measuring the effects of miRNAs there might be a difference between reporters and endogenous genes. First, even at very early time points in zebrafish embryonic development, most repression of endogenous mRNAs is attributable to miRNA-mediated deadenylation rather than direct translational repression (Subtelny et al., 2014). Second, at steady state, the fractional repression attributed to translational repression of the reporters (Béthune et al., 2012; Djuranovic et al., 2012) exceeds that typically observed for endogenous mRNAs in mammalian cells (Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010). Similarly, the magnitude of repression observed for reporters vastly exceeds that typically observed for endogenous mRNAs in mammalian cells.

Here, we substantially expand the contexts and conditions for which the repressive effects on endogenous mRNAs are examined. We measured the consequence of deleting specific miRNAs on the mRNA and translation (or protein) of predicted targets in mouse liver, primary macrophages, and activated and nonactivated primary B cells, thereby adding four additional biological settings to the previous two settings (mouse neutrophils and zebrafish embryos) in which translational effects on endogenous targets have been broadly measured. We also measured the translational effects on endogenous mRNAs after adding specific miRNAs in two additional cell lines (U2OS cells and NIH 3T3 cells) and two additional conditions (growth-arrested cells and translationally inhibited cells). In all cases, mammalian miRNAs predominantly acted to decrease target mRNA levels, with relatively small contributions from translational repression. We then examined the repression dynamics of

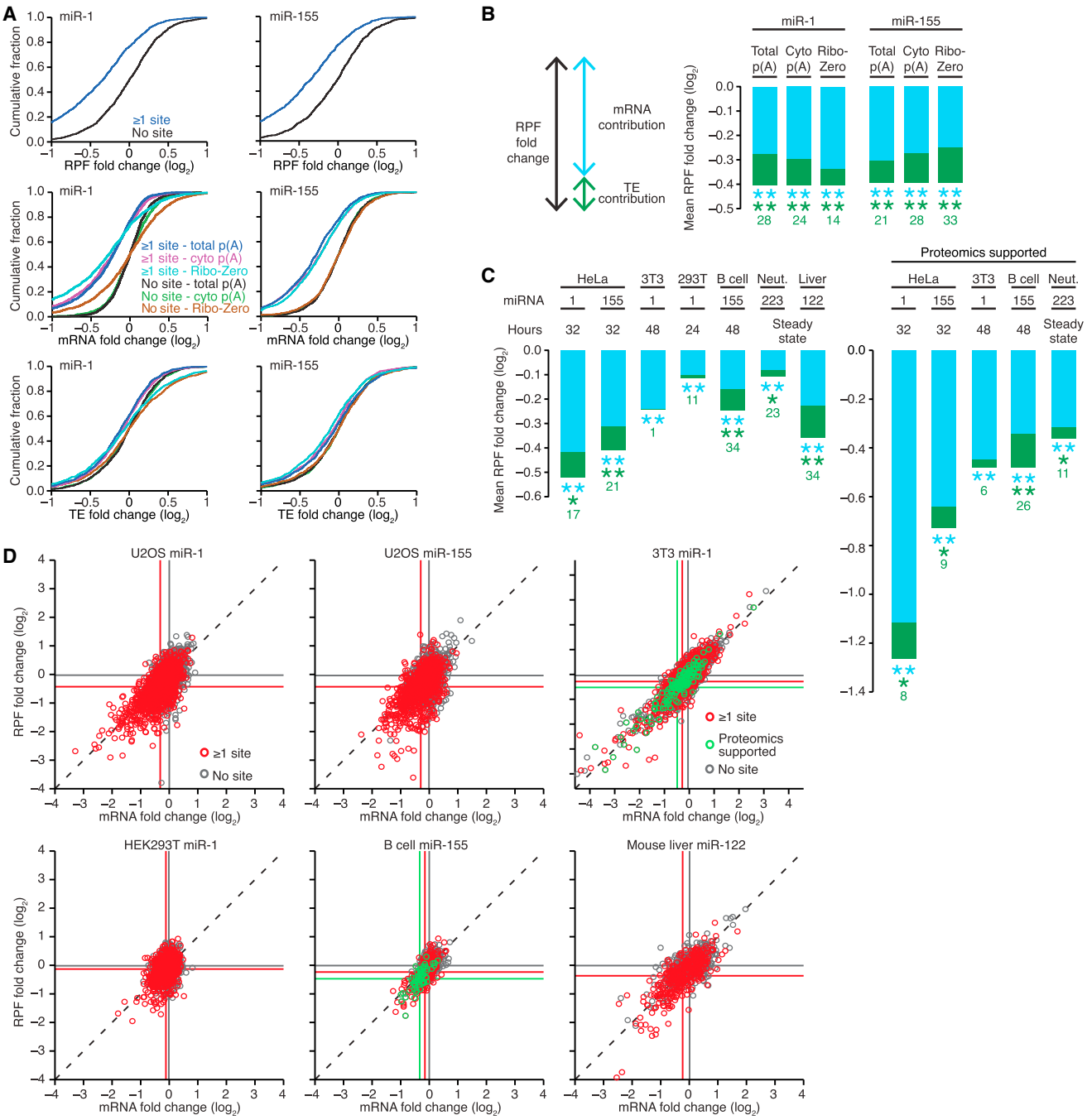


Figure 1. Steady-State Changes in Gene Expression Due to miRNAs

(A) The influence of using different types of mRNA enrichment when measuring the effects of miRNAs on mRNA levels and TE. Plots show cumulative distributions of changes in RPFs (top), mRNA (middle), and TE (bottom) after transfection of either miR-1 (left) or miR-155 (right) into U2OS cells. The impact of the miRNA on genes with at least one site to the cognate miRNA in their 3′ UTR (≥1 site; n = 1,321 and 1,075 for miR-1 and miR-155, respectively) is compared to that of control genes (no site; n = 1,205 and 1,056, respectively), which were chosen from the genes with no site to the cognate miRNA throughout their entire transcript to match the 3′ UTR length distributions of site-containing genes. The three types of mRNA enrichment were poly(A)-selected total RNA, poly(A)-selected cytoplasmic RNA, and tRNA/rRNA-depleted total RNA (total p(A), cyto p(A), and Ribo-zero, respectively). RNA-seq analyses of these preparations were used to calculate mRNA and TE changes, with results plotted as indicated in the key. Data were normalized to the median changes observed for the controls. See also Figure S1.

(B) A simplified representation of the results in (A) showing for each experiment the mean RPF fold change (log_2) attributable to changes in mRNA (blue) and TE (green), after subtracting the mean RPF change of the no-site control genes. The bars for the percent contribution attributable to mRNA and TE changes are calculated using the mean RNA and RPF fold changes (log_2) after normalizing to the median no-site fold change (log_2) (Figure S2). The schematic (left) depicts the components of the compound bar graphs (right). Significant changes for each component are indicated with asterisks of the corresponding color

(legend continued on next page)

endogenous mRNAs and did not observe an early phase in which dominant translation effects imparted substantial repression. We conclude that although translational repression is rapid, its effect is relatively weak, and thus by the time consequential repression ensues, the effect of mRNA destabilization dominates.

RESULTS

Negligible Contribution of Nuclear or Deadenylated RNA to TE Changes

The adaptation of ribosome profiling to mammalian cells has provided a sensitive and quantitative method to assess the influence of miRNAs on TE (Guo et al., 2010). Ribosome profiling uses high-throughput sequencing of ribosome-protected fragments (RPFs) to determine the positions of millions of ribosomes on mRNAs (Ingolia et al., 2009). To assess the TE of a gene, RPFs mapping to its open reading frame are normalized to its mRNA abundance, as determined by RNA sequencing (RNA-seq).

When comparing samples with and without a particular miRNA, the change in RPFs for a target of that miRNA reflects the aggregate effects of mRNA degradation and translational repression, while the change in mRNA reflects only the component attributable to degradation. After accounting for the change in RPFs attributed to mRNA degradation, the residual change in RPFs reflects a change in TE, which is interpreted as the miRNA-mediated translational repression acting on the message at the moment the ribosomes were arrested.

Previously, we observed little miRNA-mediated translational repression in mammalian cells, with the concern that these modest TE changes might actually be overestimates (Guo et al., 2010). An overestimation would occur if some polyadenylated mRNA were sequestered away from the compartment containing both miRNAs and ribosomes, as would be the case for mRNAs awaiting export from the nucleus. In this case, miRNA-mediated degradation of mRNAs only in the cytoplasm would lead to a larger relative loss of RPFs (which are only from the cytoplasm) than mRNA fragments (which are from both the nucleus and cytoplasm), thereby inflating the apparent translational repression. To address this concern, we performed ribosome profiling on miRNA- and mock-transfected U2OS cells and, in parallel, performed RNA-seq on poly(A)-selected RNA from both whole-cell lysates and cytoplasmic fractions. The efficacy of fractionation was demonstrated by the depletion of preribosomal RNAs (pre-rRNAs) in the cytoplasmic fraction (Figure S1A, available online). Following transfection of miR-1, a miRNA not normally expressed in U2OS cells, repression was observed, with significant degradation of mRNAs with at least one miR-1 3′ UTR site (Figure 1A). The amount of degradation was indistinguishable in the RNA-seq libraries made with either whole-cell or cytoplasmic mRNA, and thus the amount of translational repression was similarly indistinguishable (Figure 1A). The same was observed with miR-155, another miRNA not normally expressed in U2OS cells, demonstrating that a nuclear mRNA sequestration artifact does not detectably elevate the signal for miRNA-mediated translational repression in mammalian cells.

A second concern involved the measurement of poly(A)-selected RNA. Monitoring changes in poly(A)-selected RNA leaves unanswered the question of whether repressed mRNAs are degraded or merely deadenylated, and underrecovery of partially deadenylated messages during poly(A) selection might overestimate the amount of mRNA degradation that has occurred. To address this concern, we generated a third set of RNA-seq libraries from the aforementioned U2OS cells, starting with whole-cell RNA preparations that were not poly(A) selected and instead were depleted of both tRNAs and rRNAs. Greatly increased RNA-seq coverage of replication-dependent histone mRNAs, which lack poly(A) tails, illustrated our ability to detect RNAs regardless of poly(A) tail length (Figure S1B). Results for miRNA-dependent changes in tRNA/rRNA-depleted RNA were similar to those of poly(A)-selected RNA (Figure 1A), which indicated that changes in accumulation of mRNA refractory to poly(A) selection were negligible. These results imply that the absolute amount of deadenylated mRNAs and other intermediates underrepresented in poly(A)-selected RNA is small, even for repressed mRNAs, presumably because these decay intermediates are rapidly decapped and degraded. Thus, concerns that translational repression measurements might have been either under- or overestimates appear to be unfounded; comparing TEs calculated by simply normalizing RPF changes to those of poly(A)-selected RNA accurately measures translational repression in mammalian cells.

To aid comparisons, the results in Figure 1A can be summarized in compound bar graphs (Figure 1B). For each experiment, the mean RPF fold change (distance that the compound bar extends below zero) indicates the overall repression. The mRNA contribution (blue component of the compound bar) indicates the extent to which mRNA degradation explains this repression, and any residual RPF change is the TE contribution (green

(*$p \leq 0.05$; **$p \leq 0.001$, one-tailed Kolmogorov-Smirnov test [K–S test]), with the relative contribution of TE to repression (Figure S2D) reported as a percentage in green below each bar. See also Figure S2.

(C) The steady-state effects of miRNAs in a variety of cell types, shown using compound bar graphs like those of (B). For comparison with our current results, previously published results from HeLa and neutrophils (neut.) (Guo et al., 2010) are also plotted after reanalysis using the current methods (including the method for choosing no-site control cohorts). When available, proteomics-supported predicted targets were also analyzed (right). For HeLa and neutrophil, these were the ones selected previously (Guo et al., 2010), and for the other samples, these were selected from our proteomics data as the subset of site-containing genes with fold changes ($\log_2$) ≤ -0.3 in the presence of the miRNA. Experiments with cell lines compared cells with and without the miRNA introduced by either transfection (HeLa and 293T) or induction from a transgene (3T3). Experiments with B cells, neutrophils, and liver compared cells/tissues isolated from wild-type and miRNA knockout mice. The hours indicate the time following transfection (HeLa and 293T), induction (3T3), or activation (B cells). See also Figure S3 and Tables S1 and S2.

(D) Comparison of mRNA and RPF changes for individual genes analyzed in (A)–(C). For U2OS cells, the results for the poly(A)-selected cytoplasmic RNA are shown. The dashed line is for y = x; the vertical and horizontal lines indicate the mean fold changes for the correspondingly colored groups of genes. Red, genes with ≥ 1 3′ UTR site to the cognate miRNA; gray, no site to the miRNA selected as in (A); green, proteomics-supported predicted targets (Tables S1 and S2). Data were normalized to the median changes observed for the controls. A comparable analysis of the HeLa and neutrophil data has been published (Guo et al., 2010).

component), which reflects the translational repression of the remaining mRNA. Based on the RPF reductions attributable to these two repression modes, their relative contributions to repression are then calculated (Figure S2). Of the two modes, mRNA degradation dominates in U2OS cells (Figure 1B), despite the presence of a P body subtype reported to impart increased translational repression (Castilla-Llorente et al., 2012).

Dominant mRNA Destabilization in Many Contexts

We expanded our analysis to examine the steady-state effects of gaining or losing a miRNA in additional cell lines and biological contexts. These experiments included studies comparing RPF and mRNA measurements in liver from wild-type mice, which expresses miR-122, to those in liver from mice lacking the *mir-122* gene. Similarly, the effects of miR-155 in activated primary murine B cells were measured comparing cells from wild-type mice to those lacking the *mir-155* gene. These loss-of-function experiments enabled analysis of endogenous targets in their endogenous settings. The effects on predicted targets of endogenous miR-122 in mouse liver, endogenous miR-155 in primary mouse B cells, induced miR-1 (expressed from a transgene) in 3T3 cells, and transfected miR-1 in HEK293T cells all resembled the published effects of endogenous miR-223 in neutrophils and transfected miRNAs in HeLa cells (Figure 1C). In all settings, reduced mRNA levels explained most of the steady-state RPF reduction observed in the presence of the miRNA, implying that miRNAs predominantly act to reduce target mRNA levels. Nonetheless, mean RPF reduction attributable to translational repression was observed, ranging from 1%–34% of the total, depending on the experiment.

Because a 7–8 nt site to a miRNA is not always sufficient to mediate miRNA targeting, high-throughput proteomic measurements can be used to identify high-confidence targets by identifying site-containing genes with less protein in the presence of the miRNA (Guo et al., 2010). With this in mind, we performed a quantitative proteomics experiment using SILAC (stable isotope labeling with amino acids in culture) to identify a set of genes with reduced protein after inducing miR-1 in 3T3 cells (Table S1) and pulsed SILAC (Selbach et al., 2008) to identify those responding to miR-155 in activated B cells (Table S2). These proteomics-supported predicted targets showed greater mean repression than did the complete set of genes with $\geq$1 site, as expected if they were enriched in direct targets of the miRNA (Figure 2C). For new and published experiments with proteomics-supported predicted targets, the fractional repression attributed to translational repression ranged from 6%–26%, somewhat narrower than the range observed when considering all mRNAs with sites, perhaps because a focus on the more confidently identified targets decreased experimental variability.

Although the amount of repression attributed to translational repression did not always reach statistical significance, our results are consistent with the idea that a small amount of translational repression occurs for each direct target in each context. As was found previously (Baek et al., 2008; Hendrickson et al., 2009; Guo et al., 2010), a gene-by-gene analysis of results from each of the examined settings revealed no compelling evidence for a subset of genes repressed at only the translational level (Figure 1D), although the possibility of a few such genes cannot be ruled out.

Matching mRNA and Proteomic Results for Less-Proficient miRNAs

In pilot experiments aimed at extending our studies to other endogenous contexts, we used wild-type and miRNA-deleted mice to acquire mRNA microarray data for macrophages and neutrophils with and without miR-21 and B cells with and without miR-150. Although these miRNAs were each among the most frequently sequenced miRNAs in the respective wild-type cells (Figure S3A), we observed weak miRNA effects when comparing sets of genes with and without 3′ UTR sites to the cognate miRNA (Figure S3B).

A potential explanation for the weak signals observed by mRNA profiling was that most of the repression was occurring through translational repression rather than mRNA degradation. However, when we used quantitative proteomics to test this possibility, the proteomics results mirrored those of the mRNA arrays, providing no evidence for substantial translational repression (Figure S3B and Table S3). Thus, the weak repression signals observed at the mRNA level for endogenously expressed miR-21 and miR-150 were not due to a discrepancy between mRNA changes and the overall effects of miRNA-mediated repression. These results add to the growing list of endogenous settings for which mRNA changes accurately represent the effects of miRNA-mediated repression. This list now includes miR-223 in neutrophils (Baek et al., 2008; Guo et al., 2010), miR-21 in macrophages and neutrophils (Figure S3B), miR-122 in liver (Figure 1C), miR-150 in primary B cells (Figure S3B), and miR-155 in activated B cells (Figure 1C).

Dynamics of Endogenous mRNA Repression by Inducible miRNAs

The shifting regulatory regime in the early zebrafish embryo, which changes the consequences of miRNA-mediated poly(A) tail shortening, confounded the previous attempt to determine the dynamics of the two modes of repression for endogenous messages (Bazzini et al., 2012; Subtelny et al., 2014). Therefore, we set out to characterize the regulatory dynamics of miRNA-mediated repression of endogenous mRNAs and determine if there might be an endogenous setting in which these dynamics could give rise to a phase of substantial translation-dominated repression, as previously observed in reporter experiments (Béthune et al., 2012; Djuranovic et al., 2012).

Perhaps the most dynamic mammalian miRNA is miR-155, which is rapidly and strongly induced in B and T cells upon activation (Thai et al., 2007). In primary murine B cells, we observed a nearly 10-fold increase 4 hr after activation with lipopolysaccharide, interleukin-4 (IL-4), and anti-CD40 (Figure 2A). Although presumably not as strong as for miR-430 in zebrafish embryos (which is expressed from as many as 93 loci; Giraldez et al., 2005), miR-155 induction was nonetheless stronger than that of other mammalian miRNAs in that no other mammalian miRNA has been reported to increase so rapidly to a high level of expression.

To assess the dynamics of translational repression and mRNA decay during miR-155-mediated repression, we isolated B cells from wild-type and miR-155 knockout mice, activated these cells, and then performed ribosome profiling and RNA-seq to monitor miRNA-dependent TE and mRNA changes occurring soon after

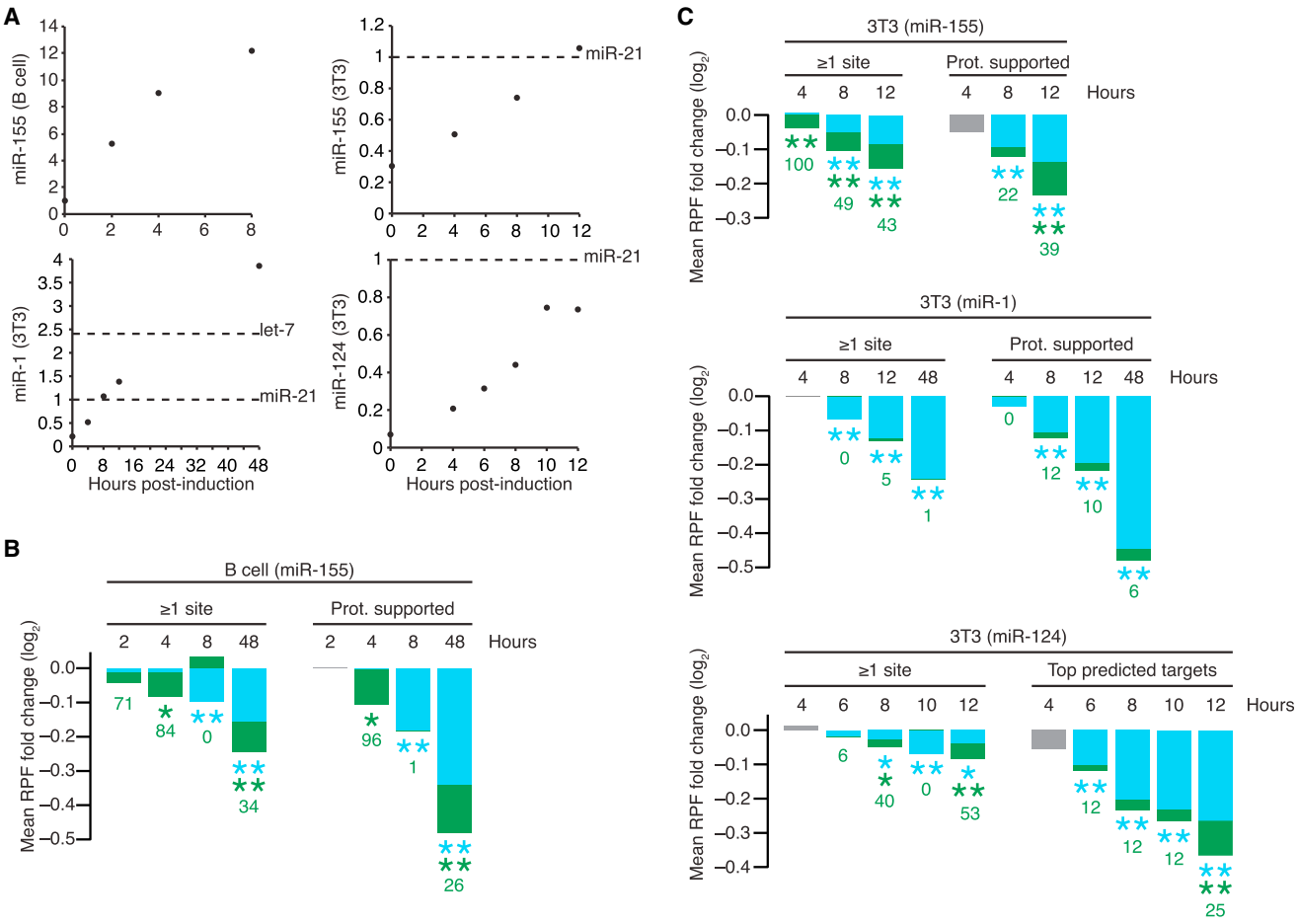


Figure 2. Minor Impact of Translational Repression at All Times in Mammalian Cells

(A) Induction of miRNAs in activated murine B cells and in contact-inhibited NIH 3T3 cells engineered to inducibly express miR-1, miR-124, or miR-155. Induction was monitored using RNA blots, probing for the induced miRNA. For samples from B cells, the membrane was reprobed for endogenous U6 snRNA, which served as a loading control for normalization, and expression is plotted relative to that of the nonactivated cells. For samples from 3T3 cells, synthetic standards for the induced miRNAs and endogenous miR-21 were included on the blot and used for absolute quantification. Expression is plotted relative to that of miR-21, with relative expression of the let-7 family (inferred from small-RNA sequencing data) also shown.

(B) The contributions of mRNA decay and translational repression following miR-155 induction in primary murine B cells. The same sets of site-containing and control genes are analyzed in all time points. If the contribution of TE was calculated to be less than 0, the value reported below the bar was 0; otherwise, as in Figure 1C. The 48 hr time point is replotted from Figure 1C and was from a preparation of cultured B cells independent from that used for the earlier time points. See also Table S2.

(C) The contributions of mRNA decay and translational repression following induction of miR-155 (top), miR-1 (middle), or miR-124 (bottom) in the corresponding contact-inhibited 3T3 cell lines. In the absence of proteomics data for miR-124, the top 100 site-containing genes, as ranked by total context+ score (Garcia et al., 2011) regardless of site conservation, were analyzed to focus on a subset of site-containing genes likely to be regulated by miR-124; otherwise, as in (B). The miR-1 48 hr time point is replotted from Figure 1C and is from the same experiment as the earlier time points. See also Tables S1 and S2.

induction. At 2 hr postactivation, repression of genes with ≥1 site was detectable, but neither the mRNA nor the TE component was significantly decreased on its own. At 4 hr, the small amount of repression was predominantly attributable to reduced TE (Figure 2B). By 8 hr, the proportion attributed to translational repression abated, and at this time point the mRNA degradation so closely approached overall repression that the mean mRNA change for genes with ≥1 site slightly exceeded the mean RPF change (Figure 2B; $p = 0.028$, two-tailed K–S test for TE). Because this slight excess was not observed for the proteomics-supported predicted targets (Figure 2B) or in similar experiments with other miRNAs, we attribute it to experimental variability rather than translational activation. After 48 hr, mRNA degradation continued to dominate (Figure 2B; as already shown in the steady-state analyses of Figure 1C), which indicated that B cells resemble other cells with respect to steady-state repression.

Although we found some evidence for translational repression dominating early in miR-155 induction, the amount of repression observed during this brief period was much weaker than that observed during the analogous phase of reporter experiments. Thus, we cannot claim to have found a mammalian setting with an early phase of substantial translational repression of

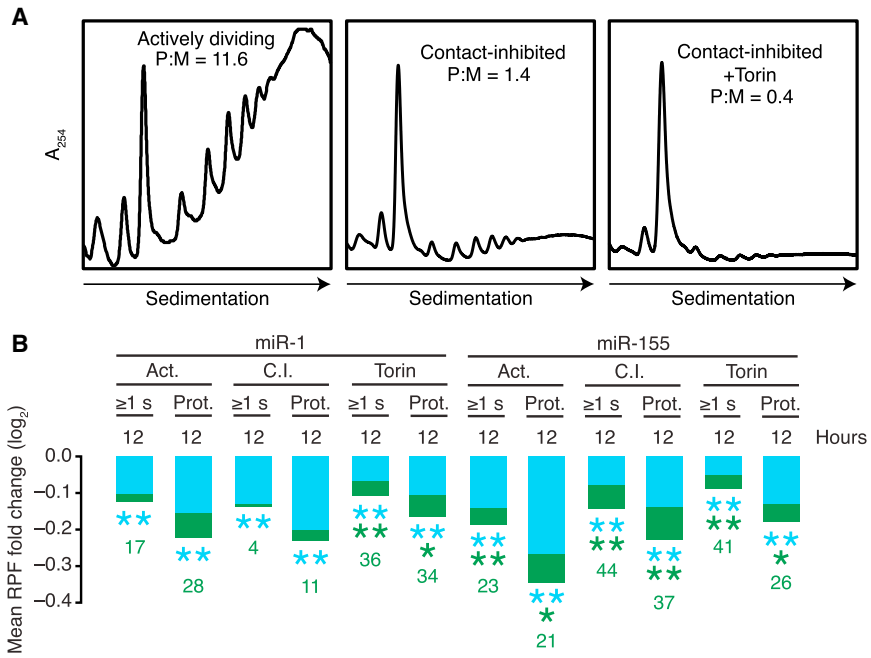


Figure 3. Negligible Influence of Translational Stress or State on the Repression Mode

(A) Polysome profiles showing the translational activity of actively dividing (left), contact-inhibited (middle), and Torin1-treated contact-inhibited (right) miR-1 inducible 3T3 cells. Profiles are normalized to the monosome peak, with the polysome-to-monosome ratio (P:M) indicated.

(B) The contributions of mRNA decay and translational repression following miR-1 or miR-155 induction in the corresponding 3T3 cell lines in the indicated states; otherwise, as in Figure 2C. Results for contact-inhibited 3T3 cells expressing miR-1 and miR-155 were recalculated so as to only consider site-containing and no-site genes present in all samples. Act., actively dividing; C.I., contact-inhibited; Torin, contact-inhibited and Torin1-treated; ≥1 s, genes with at least one site to the cognate miRNA in their 3′ UTR; Prot., proteomics-supported predicted targets. See also Figure S4.

endogenous messages, i.e., a time at which substantial repression would be missed if only mRNA changes were monitored. To further explore repression dynamics in mammalian cells, we created stable, miRNA-inducible 3T3 cell lines in which doxycycline treatment rapidly induced the expression of a miRNA not normally expressed in 3T3 cells (either miR-1, miR-124, or miR-155) to levels comparable to those of miR-21 and the let-7 miRNA family (Figure 2A), which are the miRNA and miRNA family most frequently sequenced for these cells (Rissland et al., 2011). The major advantage of such cell lines for studying the dynamics of translational repression and decay on endogenous messages is that, in contrast to B cells, miRNA induction does not accompany significant developmental changes, allowing the miRNA effects to be more easily isolated. With these lines, we performed ribosome profiling and RNA-seq soon after miRNA induction, comparing translational efficiencies and mRNA expression levels with those of uninduced cells.

To account for the 2–3 hr lag prior to the appearance of increased mature miRNA, the first time point examined was 4 hr postinduction. At 4 hr, the miR-155-expressing line showed significant repression of genes with ≥1 site, all of which was attributed to translational repression (Figure 2C). At later time points, mRNA degradation dominated, as observed in B cells. For the miR-1-expressing line, 4 hr was too early to observe significant repression for genes with ≥1 site, and by 8 hr, mRNA degradation already dominated (Figure 2C), suggesting that we had missed any potential translation-dominant phase. For miR-124, a translation-dominant phase also was not observed (Figure 2C), presumably because induction was too gradual to achieve significant repression at early time points (as we did not acquire murine proteomics data for miR-124, the top predicted targets were used instead of proteomics-supported predicted targets). Because miRNA induction in vivo is rarely more rapid than that achieved for miR-124 in our inducible line, we suggest that the miR-124 results are representative of most endogenous settings.

Minimal Influence of Translational Stress and State

Having investigated eight different cell types and six different miRNAs, and having considered both pre-steady-state and later time points without identifying a setting with substantial overall repression in which translational effects dominated, we turned to the potential influence of cellular state. Studies of *lin-4*-mediated repression in *C. elegans* suggest that starvation might tip the balance toward more translational inhibition with less mRNA degradation (Holtz and Pasquinelli, 2009), presumably because starvation influences global translational activity. Therefore, we compared the relative contributions of TE and mRNA degradation for 3T3 cells in three translational states: (1) dividing cells, which have very active translation (polysome to monosome ratio [P:M] = 11.6), (2) contact-inhibited cells (P:M = 1.4), and (3) contact-inhibited cells under Torin1-induced mammalian target of rapamycin (mTOR) inhibition (P:M = 0.4) (Figure 3A). We found no pervasive difference in the relative contribution of translational repression to miR-1- and miR-155-mediated repression between these states (Figure 3B), despite the ~30-fold range in translational activity. Thus, translational stress, and more generally the translational state, does not have a perceptible global impact on the mode of miRNA-mediated regulation in these mammalian cells.

Because translating ribosomes displace miRNA-directed silencing complexes, which renders miRNA sites in the path of the ribosome much less effective than those ≥15 nt downstream of the stop codon (Grimson et al., 2007), we reasoned that the efficacy of sites within open reading frames (ORFs) might increase in conditions of reduced translational activity. Indeed, relative to the efficacy of 3′ UTR sites, the efficacy of ORF sites did appear to increase when translation was repressed with Torin1

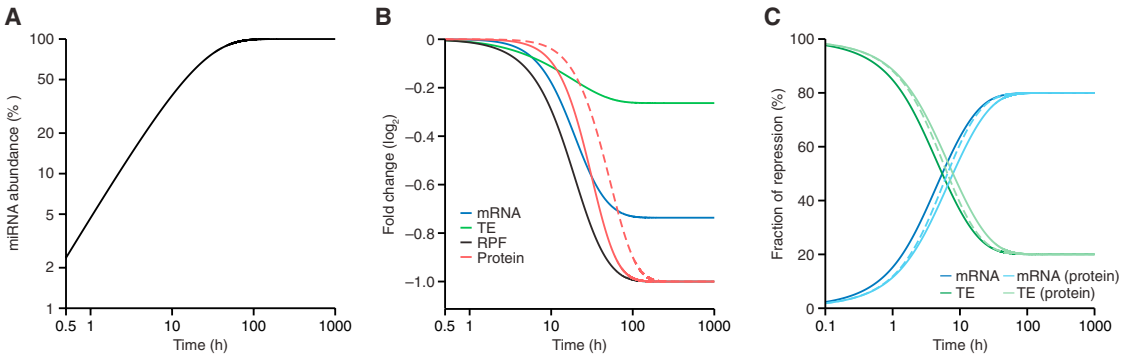


Figure 4. Simulated Dynamics of miRNA-Mediated Repression

(A) Simulation of rapid miRNA induction that begins with no miRNA and rises to a concentration exceeding that of the highest expressed endogenous miRNA in 3T3 cells within 6 hr.

(B) Changes in target mRNA (blue), TE (green), RPF (black), and protein (red; solid line, 10 hr protein half-life; dashed line, 100 hr protein half-life) levels resulting from the miRNA induction in (A).

(C) The relative contributions of mRNA decay and translational repression to the overall repression in (B) when measured at either the RPF level (dark blue and dark green, respectively) or the protein level (light blue and light green; solid lines, 10 hr protein half-life; dashed lines, 1 hr protein half-life).

(Figure S4A), which supported the model in which displacement of bound miRNAs by translating ribosomes is the predominant reason that ORF sites are ineffective.

DISCUSSION

The Principles of Repression Dynamics

Our results in 3T3 and B cells, considered in light of the fundamental differences between the nature of translational repression and mRNA destabilization, lead to the following principles regarding the miRNA-mediated repression of endogenous mRNAs in mammalian cells: compared to translational repression, detectable mRNA destabilization occurs after more of a lag, presumably because mRNA decay takes longer than inhibiting translation initiation. Because of this relative lag, after unusually robust miRNA induction, we can detect a short phase resembling that observed in reporter experiments, in which most of the repression is from decreased TE. However, the lag in destabilization does not last long, and destabilization soon dominates. To illustrate these principles, we simulated the repression time course of a rapidly induced miRNA for which 80% of the steady-state repression is through mRNA destabilization and 20% is through translational repression (Figure 4). In our simulation, translational repression begins immediately upon miRNA-mRNA association, and mRNA degradation occurs through an increased degradation rate for the miRNA-bound mRNA. This approach yields an early phase in which translational repression dominates, consistent with that observed in our experimental time courses (Figure 4B). The transition from mostly translational repression to mostly mRNA destabilization takes place at 5.7 hr (Figure 4C), when relatively little overall repression (9.7% RPF decrease, compared to a 50% decrease at steady state) is occurring (Figure 4B). Our example simulates very rapid miRNA induction; within 6 hr the induced miRNA reaches levels that would make it the highest expressed miRNA in 3T3 cells (Figure 4A), similar to or faster than the induction observed in our 3T3 cell lines (Figure 2A). Slowing the induction rate by about half would result in this transition occurring at a point of even less repression (6.6% RPF decrease), and thus in most mammalian contexts miRNA induction would be too slow to yield detectable repression during the phase in which TE changes dominate. For an early phase of substantial repression mediated primarily through TE changes, miRNA induction would have to be stronger than that ever reported, which is consistent with our inability to find a mammalian context with substantial translation-based repression.

Decreases in mRNA and TE lead to decreased protein from the targeted messages, and this change in protein is what matters to the cell. Despite the ultimate importance of the protein changes, measuring these changes over time is less informative for analyzing miRNA repression dynamics than is measuring RPF and mRNA changes, which more directly captures the molecular effects of the miRNA in inhibiting translation and destabilizing mRNA. RPF and mRNA measurements are also more suitable for quantitative comparisons for two reasons: (1) they enabled accurate comparisons of more miRNA targets and (2) they were each acquired using analogous methods that measured differences at one moment in time without the complications that arise from pre-steady-state measurements of protein changes. With regard to these complications, protein differences detected using direct labeling or standard metabolic labeling (e.g., SILAC) cannot distinguish between protein synthesized before or after induction of the miRNA and thus are unsuitable for pre-steady-state measurements because they would underestimate the impact on newly synthesized protein. Pulsed SILAC differentiates between preexisting and newly synthesized protein but as currently implemented still entails an extended period ($\geq$6 hr for global measurements) of metabolic labeling (Schwanhäusser et al., 2009; Huo et al., 2012), which compromises its utility for observing the results of the first few hours of repression.

Despite the advantages of measuring RPF and mRNA changes, we note that during pre-steady-state conditions the relative TE and mRNA effects can underestimate the relative contribution of translational repression to miRNA-mediated

repression at the protein level. For example, at a given time point reduced mRNA might explain 80% of the RPF effect, leaving only 20% of the reduced protein synthesis at that moment to be explained by translational repression, but when considering the reduced protein levels (not current protein synthesis) more repression might have been due to translational repression. This is because the reduced protein levels are a function of the miRNA effects integrated up to the current time point, which includes earlier periods in which translational repression might have represented a greater share of the decreased protein synthesis.

The extent to which the relative contribution of translational repression would be underestimated depends on three factors: (1) the extent to which translational repression represents a greater share of the overall repression at the earlier time periods, (2) the relative strength of the overall repression during earlier periods, and (3) the stability of the protein. Our results indicate that with respect to the second factor, the relative strength of the overall repression during earlier periods is low in mammalian contexts, which implies that any underestimate of the contribution of translational repression to the reduction in protein levels would be minimal. In our simulation, the greatest underestimate was observed at 5.7 hr, when TE changes explained 49% of the reduction in protein synthesis at that moment and 58% of the reduction in protein accumulation, assuming intermediate protein stability (10 hr protein half-life; Figure 4C). A shorter protein half-life further diminished the small differential between protein synthesis and protein accumulation (Figure 4C), whereas a longer half-life delayed the onset of any consequential miRNA effect on protein abundance to a period well beyond the onset of substantial mRNA decay (Figure 4B). In sum, monitoring protein levels rather than TE would not increase the prospects for finding a mammalian setting in which substantial translational repression dominates.

Comparison of Fish Embryos and Mammalian Contexts

Attempts to characterize the dynamics of the two modes of miRNA-mediated repression in zebrafish embryos were confounded by two unique features of fish and frog embryos prior to gastrulation: (1) a strong coupling between poly(A) tail length and translational efficiency and (2) an unusual mRNA metabolism wherein mRNAs with short poly(A) tails are stable. These features do not necessarily preclude analysis of dynamics, but in these contexts changes in TE due to miRNA-mediated deadenylation must be accounted for independently of changes in TE due to direct miRNA-mediated translational repression. Indeed, when the repression due to mRNA decay is thought of as including deadenylation-dependent translational repression, mRNA decay is the predominant mode of miRNA-mediated repression at all time points analyzed in zebrafish (Subtelny et al., 2014) just as it is at all but the earliest time points in mammalian cells. An important difference between most mammalian systems and early developmental systems (and presumably neuronal synapses or other systems with the aforementioned features) is that, in the latter, effects on translation must be measured to accurately capture the impact of the miRNA on gene expression, and effects on deadenylation must be measured to understand how repression is achieved. However, neither system seems to have a phase in which deadenylation-independent translational repression performs substantial repression without even stronger repression detectable by mRNA changes.

Mechanistic Interpretations

Although translational repression and mRNA decay both lead to reduced protein synthesis, the mechanism used for repression has important biological implications. To the extent that repression occurs through translational repression, rapid recovery would be possible without requiring new transcription. This would, for example, be the case in early zebrafish embryos, where the repression of miRNA targets could be rapidly reversed through cytoplasmic polyadenylation. In most settings, however, reversal of miRNA-mediated repression requires new transcription, as mRNA decay constitutes the major mode of repression.

When miRNA-mediated mRNA decay was first reported, it was proposed to occur either through active recruitment of mRNA degradation machinery or as a secondary effect of inhibiting translation (Lim et al., 2005). Although we observe translational repression prior to the decay of endogenous mRNAs in some experiments, this temporal relationship does not imply that mRNA decay is a consequence of translational repression because it is also consistent with mRNA decay simply being a slower process. Indeed, several observations favor the model that the decay occurs through active recruitment of mRNA degradation machinery rather than as a secondary effect of inhibiting translation. First, miRNA targeting can destabilize reporter transcripts that cannot be translated, which indicates that mRNA destabilization is not merely a secondary effect of reducing the number of ribosomes translating an mRNA (Mishima et al., 2006; Wu et al., 2006; Eulalio et al., 2007; Wakiyama et al., 2007; Eulalio et al., 2009; Fabian et al., 2009), although it does not rule out models in which only translationally repressed mRNAs can be destabilized. Second, direct biochemical interactions link miRNAs to Argonaute, Argonaute to TNRC6, and TNRC6 to the deadenylase complexes (the PAN2-PAN3 complex and the CCR4-NOT complex) that shorten the poly(A) tail (Braun et al., 2012), thereby showing how the mRNA degradation machinery can be actively recruited independent of either the act or the consequence of translational repression. Finally, our work greatly expands the number of mammalian systems examined and shows that in each of these systems mRNA destabilization explains a large majority (from 66%–>90%) of the miRNA-mediated repression observed at steady state.

The idea that the mRNA destabilization might be a secondary consequence of inhibiting translation would be more plausible if a larger fraction of the steady-state repression was through translational repression; otherwise, the mRNA destabilization is out of proportion to the translational repression. We are not aware of any mammalian examples in which translationally repressed messages are so destabilized as a secondary consequence of this repression that the amount of steady-state destabilization exceeds the amount of steady-state translational repression. Indeed, the idea that mammalian messages might be destabilized solely as a secondary consequence of reduced ribosome occupancy or density appears to be largely an extrapolation from observations made in bacteria and yeast, but not mammalian cells (Muhlrad et al., 1995; Schwartz and Parker, 1999; Deana and

Belasco, 2005). When examining mammalian mRNAs in general (irrespective of miRNA targeting), we find only a very weak correlation between TE and mRNA half-life (Figure S4B, R^2 = 0.004 and 0.001 for 3T3 and HeLa, respectively), and others have shown that repression of translational initiation through the iron response element (a textbook example of mammalian translational repression) does not impart detectable destabilization of either its endogenous host mRNAs (Coccia et al., 1992; Melefors et al., 1993; Kim et al., 1996) or a reporter transcript (Hentze et al., 1987). Thus, when considered together, the available evidence strongly supports a model in which miRNAs actively recruit the deadenylation machinery, and the ensuing deadenylation, decapping, and decay comprises the major mode of miRNA-mediated repression of endogenous targets in mammalian cells.

Some translational repression accompanies mRNA destabilization as a minor component of endogenous target repression in mammalian cells. Like mRNA destabilization, this translational repression also appears to depend on recruitment of CCR4-NOT, but three observations indicate that this repression is not simply a consequence of shortened poly(A) tails. First, mRNAs without poly(A) tails can be translationally repressed (Wu et al., 2006; Eulalio et al., 2008, 2009; Braun et al., 2011; Chekulaeva et al., 2011; Zekri et al., 2013). Second, mutant complexes lacking deadenylase activity can nonetheless promote translational repression (Cooke et al., 2010). Third, tail length and TE are not correlated in most mammalian settings (Subtelny et al., 2014). Thus, the two modes of miRNA-mediated repression seem to represent two independent ramifications of recruiting the deadenylation complexes.

Reconciling Results with Single-Gene Studies of mRNA and Protein Changes

The conclusion that mRNA destabilization is the major mode of miRNA-mediated repression agrees with many previous observations monitoring protein and mRNA changes of single target genes after perturbing a miRNA. Among the >30,000 research studies of mammalian miRNAs, there are also counter examples in which single-gene measurements seem to suggest a greater role for translational repression (Poy et al., 2004; O'Donnell et al., 2005; Zhao et al., 2005; Chen et al., 2006). An advantage of our approach is that we simultaneously examine thousands of genes, comparing the changes of both mRNA level and TE for hundreds of genes that have at least one miRNA site to those of hundreds of genes that lack a site and thus serve as internal controls. The aggregate result of this global approach should reflect the overall contributions of mRNA destabilization and translational repression, whereas a single-gene study might choose a nonrepresentative example and reach a conclusion that does not apply more generally to the targets of the miRNA.

This raises the question as to what might explain a single-gene result in which a miRNA-dependent change is observed in protein (i.e., with an immunoblot) but not mRNA (e.g., with quantitative RT-PCR), which would appear as an outlier in our analyses. Might such outliers represent targets that are repressed at the level of translation without being destabilized? Although this possibility cannot be excluded, changes observed among our control genes that lack miRNA sites raise doubts about its validity. In most experiments (the possible exception being U2OS cells transfected with miR-155), a similar number of these control genes also change at the level of translation without being destabilized (Figure 1D). The observation that this behavior usually does not depend on the presence of a site to the miRNA suggests that either indirect effects of the miRNA or experimental variability explain the presence of most outliers that appear to be changing only at the level of translation.

Other single-gene examples for which translational repression is reported to be the major mode of miRNA-mediated regulation examine reporter mRNAs rather than endogenous mRNAs (Doench and Sharp, 2004; Kiriakidou et al., 2004; Nelson et al., 2004; Yekta et al., 2004; Pillai et al., 2005). Interestingly, the fractional component of regulation attributable to translational repression generally seems to be higher for reporters than for endogenous genes. We have begun experiments that aim to understand this difference between reporter and endogenous genes. Once this difference is understood, reporters could be developed that better recapitulate the regulation of endogenous genes, which would provide more relevant tools for studying the mechanism and dynamics of miRNA-mediated repression.

EXPERIMENTAL PROCEDURES

RNA Isolation

For RNA-seq, total RNA was extracted from B cells and U2OS cells using TRI reagent. Using TRI reagent, cytoplasmic RNA was extracted from cytoplasmic fractions of U2OS cells that were separated from nuclear fractions by differential centrifugation. Briefly, whole-cell lysate prepared as described (Guo et al., 2010) was centrifuged at 1,300 × *g* for 10 min, and the resulting supernatant was collected as the cytoplasmic fraction while the pellet obtained was collected as the nuclear fraction. To prepare rRNA/tRNA-depleted U2OS total RNA, total RNA was first treated with the Ribo-Zero rRNA removal kit (Epicenter BioTechnologies) according to manufacturer's instructions. The resulting rRNA-depleted RNA sample was then spin-filtered using Ultra-4 centrifugal filters with Ultracel-100 membranes (Amicon) by centrifuging at 5,000 × *g* for 10 min at 4°C. The filtrate was enriched in tRNAs and was discarded, and the retentate was collected as the rRNA/tRNA-depleted RNA sample. RNA for all other RNA-seq samples was prepared by extracting RNA from ribosome profiling lysates with TRI reagent as described (Subtelny et al., 2014). Except in the case of the tRNA/rRNA-depleted U2OS RNA sample, the extracted RNA was poly(A) selected as described (Subtelny et al., 2014). All animal experiments were performed in accordance with protocols approved by the MIT and Ohio State University Committees on Animal Care.

Ribosome Footprint Profiling and RNA-Seq

For B cell and U2OS samples, ribosome profiling and RNA-seq were performed essentially as described (Guo et al., 2010), with the only difference being how the RNA was isolated or enriched the cases of U2OS cytoplasmically enriched RNA and tRNA/rRNA-depleted total RNA. All other samples were prepared as described (Subtelny et al., 2014). Detailed protocols are available at http://bartellab.wi.mit.edu/protocols.html. Reference transcript annotations were downloaded (in refFlat format) from the UCSC Genome browser, and for each gene the longest transcript was chosen as a representative transcript model. RPF and RNA-seq reads were mapped to ORFs as described, which excluded the first 50 nt of each ORF so as to eliminate signal from ribosomes that initiated after adding cycloheximide (Subtelny et al., 2014).

ACCESSION NUMBERS

The NCBI GEO accession number for all microarray and sRNA-seq data and most ribosome profiling and RNA-seq data is GSE61073. The accession number for HeLa and miR-223 neutrophil data analyzed in this study is GSE22004. The accession number for the U2OS ribosome profiling data and RNA-seq

data from poly(A)-selected total RNA and tRNA/rRNA-depleted total RNA is GSE51584. The accession numbers for HEK293T mock-treated RNA-seq and ribosome profiling data are GSM1276541 and GSM1276542, respectively. The accession numbers for the uninduced miR-155 actively dividing 3T3 RNA-seq and ribosome profiling data are GSM1276543 and GSM1276544, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.08.028.

AUTHOR CONTRIBUTIONS

S.W.E., H.G., and D.P.B. designed the study. S.W.E. and H.G. performed ribosome profiling and RNA-seq, and S.W.E. did the associated analyses. S.E.M. did the mathematical modeling. R.A.R-.M. and J.V. performed the proteomics. C.S. cultured primary cells. D.B. analyzed microarray and proteomics data. S.-h.H. and K.G. harvested liver tissue. S.W.E., H.G., S.E.M., and D.P.B. wrote the paper, with input from the other authors.

ACKNOWLEDGMENTS

We thank V. Agarwal and V. Auyeung for helpful discussions; D. Patrick, E. van Rooji, and E. Olson for miR-21 knockout mice and wild-type controls; and the Whitehead Genome Technology Core for sequencing. This work was supported by NIH grants R01GM067031 (D.P.B.) and R01CA193244 (K.G.). H.G. was supported by the Agency for Science, Technology and Research, Singapore. D.P.B. is an investigator of the Howard Hughes Medical Institute.

Received: July 25, 2014
Revised: August 21, 2014
Accepted: August 22, 2014
Published: September 25, 2014

REFERENCES

Baek, D., Villén, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71.

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by *let-7* and *lin-4* miRNAs results in target mRNA degradation. Cell *122*, 553–563.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science *336*, 233–237.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. *20*, 1885–1898.

Béthune, J., Artus-Revel, C.G., and Filipowicz, W. (2012). Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. EMBO Rep. *13*, 716–723.

Braun, J.E., Huntzinger, E., Fauser, M., and Izaurralde, E. (2011). GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. Mol. Cell *44*, 120–133.

Braun, J.E., Huntzinger, E., and Izaurralde, E. (2012). A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. Cold Spring Harb. Perspect. Biol. *4*, 4.

Castilla-Llorente, V., Spraggon, L., Okamura, M., Naseeruddin, S., Adamow, M., Qamar, S., and Liu, J. (2012). Mammalian GW220/TNGW1 is essential for the formation of GW/P bodies containing miRISC. J. Cell Biol. *198*, 529–544.

Chekulaeva, M., Mathys, H., Zipprich, J.T., Attig, J., Colic, M., Parker, R., and Filipowicz, W. (2011). miRNA repression involves GW182-mediated recruitment of CCR4-NOT through conserved W-containing motifs. Nat. Struct. Mol. Biol. *18*, 1218–1226.

Chen, J.F., Mandel, E.M., Thomson, J.M., Wu, Q., Callis, T.E., Hammond, S.M., Conlon, F.L., and Wang, D.Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. Nat. Genet. *38*, 228–233.

Chen, C.Y., Zheng, D., Xia, Z., and Shyu, A.B. (2009). Ago-TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. Nat. Struct. Mol. Biol. *16*, 1160–1166.

Coccia, E.M., Profita, V., Fiorucci, G., Romeo, G., Affabris, E., Testa, U., Hentze, M.W., and Battistini, A. (1992). Modulation of ferritin H-chain expression in Friend erythroleukemia cells: transcriptional and translational regulation by hemin. Mol. Cell. Biol. *12*, 3015–3022.

Cooke, A., Prigge, A., and Wickens, M. (2010). Translational repression by deadenylases. J. Biol. Chem. *285*, 28506–28513.

Deana, A., and Belasco, J.G. (2005). Lost in translation: the influence of ribosomes on bacterial mRNA decay. Genes Dev. *19*, 2526–2533.

Djuranovic, S., Nahvi, A., and Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. Science *336*, 237–240.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes Dev. *18*, 504–511.

Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S.F., Doerks, T., Dorner, S., Bork, P., Boutros, M., and Izaurralde, E. (2007). Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. Genes Dev. *21*, 2558–2570.

Eulalio, A., Huntzinger, E., and Izaurralde, E. (2008). GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. Nat. Struct. Mol. Biol. *15*, 346–353.

Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M., and Izaurralde, E. (2009). Deadenylation is a widespread effect of miRNA regulation. RNA *15*, 21–32.

Fabian, M.R., Mathonnet, G., Sundermeier, T., Mathys, H., Zipprich, J.T., Svitkin, Y.V., Rivas, F., Jinek, M., Wohlschlegel, J., Doudna, J.A., et al. (2009). Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. Mol. Cell *35*, 868–880.

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

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs. Nat. Struct. Mol. Biol. *18*, 1139–1146.

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

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75–79.

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

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Hendrickson, D.G., Hogan, D.J., McCullough, H.L., Myers, J.W., Herschlag, D., Ferrell, J.E., and Brown, P.O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS Biol. *7*, e1000238.

Hentze, M.W., Rouault, T.A., Caughman, S.W., Dancis, A., Harford, J.B., and Klausner, R.D. (1987). A cis-acting element is necessary and sufficient for

translational regulation of human ferritin expression in response to iron. Proc. Natl. Acad. Sci. USA *84*, 6730–6734.

Holtz, J., and Pasquinelli, A.E. (2009). Uncoupling of *lin-14* mRNA and protein repression by nutrient deprivation in *Caenorhabditis elegans*. RNA *15*, 400–405.

Huo, Y., Iadevaia, V., Yao, Z., Kelly, I., Cosulich, S., Guichard, S., Foster, L.J., and Proud, C.G. (2012). Stable isotope-labelling analysis of the impact of inhibition of the mammalian target of rapamycin on protein synthesis. Biochem. J. *444*, 141–151.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Kim, H.Y., LaVaute, T., Iwai, K., Klausner, R.D., and Rouault, T.A. (1996). Identification of a conserved and functional iron-responsive element in the 5′-untranslated region of mammalian mitochondrial aconitase. J. Biol. Chem. *271*, 24226–24230.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Genes Dev. *18*, 1165–1178.

Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K.G., Tuschl, T., Manoharan, M., and Stoffel, M. (2005). Silencing of microRNAs in vivo with 'antagomirs'. Nature *438*, 685–689.

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

Melefors, O., Goossen, B., Johansson, H.E., Stripecke, R., Gray, N.K., and Hentze, M.W. (1993). Translational control of 5-aminolevulinate synthase mRNA by iron-responsive elements in erythroid cells. J. Biol. Chem. *268*, 5974–5978.

Mishima, Y., Giraldez, A.J., Takeda, Y., Fujiwara, T., Sakamoto, H., Schier, A.F., and Inoue, K. (2006). Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. Curr. Biol. *16*, 2135–2142.

Muhlrad, D., Decker, C.J., and Parker, R. (1995). Turnover mechanisms of the stable yeast *PGK1* mRNA. Mol. Cell. Biol. *15*, 2145–2156.

Nelson, P.T., Hatzigeorgiou, A.G., and Mourelatos, Z. (2004). miRNP:mRNA association in polyribosomes in a human neuronal cell line. RNA *10*, 387–394.

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

Olsen, P.H., and Ambros, V. (1999). The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol. *216*, 671–680.

Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E., and Filipowicz, W. (2005). Inhibition of translational initiation by *Let-7* MicroRNA in human cells. Science *309*, 1573–1576.

Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., and Stoffel, M. (2004). A pancreatic islet-specific microRNA regulates insulin secretion. Nature *432*, 226–230.

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D., and Izaurralde, E. (2005). A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA *11*, 1640–1647.

Rissland, O.S., Hong, S.J., and Bartel, D.P. (2011). MicroRNA destabilization enables dynamic regulation of the miR-16 family in response to cell-cycle changes. Mol. Cell *43*, 993–1004.

Schwanhäusser, B., Gossen, M., Dittmar, G., and Selbach, M. (2009). Global analysis of cellular protein translation by pulsed SILAC. Proteomics *9*, 205–209.

Schwartz, D.C., and Parker, R. (1999). Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in *Saccharomyces cerevisiae*. Mol. Cell. Biol. *19*, 5247–5256.

Seggerson, K., Tang, L., and Moss, E.G. (2002). Two genetic circuits repress the *Caenorhabditis elegans* heterochronic gene *lin-28* after translation initiation. Dev. Biol. *243*, 215–225.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

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

Thai, T.H., Calado, D.P., Casola, S., Ansel, K.M., Xiao, C., Xue, Y., Murphy, A., Frendewey, D., Valenzuela, D., Kutok, J.L., et al. (2007). Regulation of the germinal center response by microRNA-155. Science *316*, 604–608.

Wakiyama, M., Takimoto, K., Ohara, O., and Yokoyama, S. (2007). Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. Genes Dev. *21*, 1857–1862.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. Cell *75*, 855–862.

Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. Proc. Natl. Acad. Sci. USA *103*, 4034–4039.

Yekta, S., Shih, I.H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of *HOXB8* mRNA. Science *304*, 594–596.

Zekri, L., Kuzuoğlu-Öztürk, D., and Izaurralde, E. (2013). GW182 proteins cause PABP dissociation from silenced miRNA targets in the absence of deadenylation. EMBO J. *32*, 1052–1065.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. Nature *436*, 214–220.

Appendix D.

Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression

Rémy Denzler[1], Sean E. McGeary[2,3,4], Alexandra C. Title[1], Vikram Agarwal[2,3,4,5], David P. Bartel[2,3,4], and Markus Stoffel[1,6]

[1]Institute of Molecular Health Sciences, Swiss Federal Institute of Technology in Zurich (ETH Zurich), Otto-Stern-Weg 7, 8093 Zürich, Switzerland
[2]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[3]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[4]Department of Biology
[5]Computational and Systems Biology Program
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

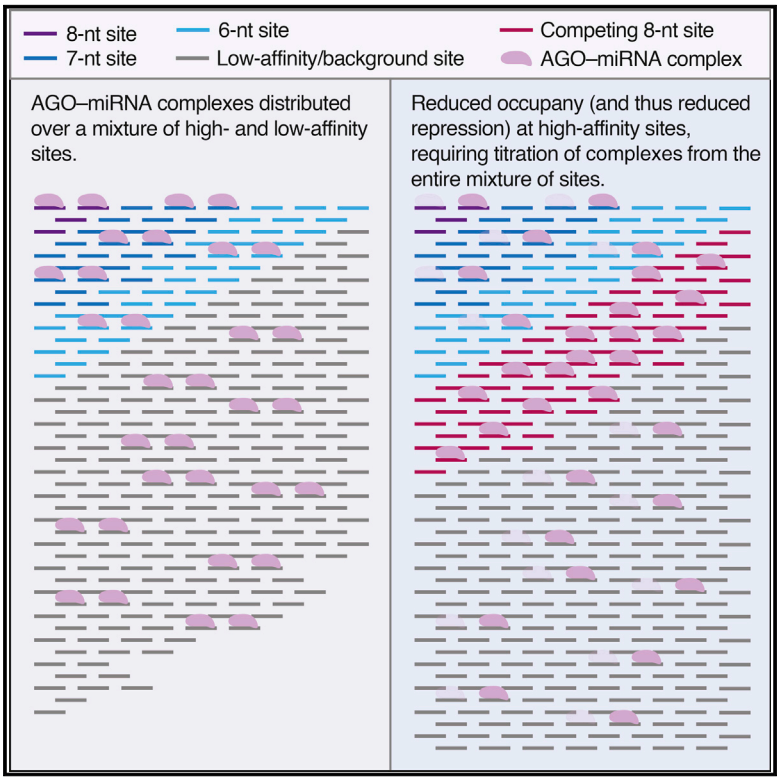
Published as:
Denzler, R., McGeary, S.E., Title, A.C., Agarwal, V., Bartel., D.P., and Stoffel, M. (2016). Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. Mol. Cell *64*, 565–579.

Article

Molecular Cell

Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression

Graphical Abstract



Authors

Rémy Denzler, Sean E. McGeary, Alexandra C. Title, Vikram Agarwal, David P. Bartel, Markus Stoffel

Correspondence

dbartel@wi.mit.edu (D.P.B.), stoffel@biol.ethz.ch (M.S.)

In Brief

Denzler et al. show that effects of competing miRNA sites are insensitive to reduced miRNA activity, low-affinity/background miRNA sites contribute to competition, and adjacent miRNA sites can cooperatively sequester miRNAs. Overall, their results reduce the prospects of observing an effect from a ceRNA.

Highlights

- ceRNA-mediated derepression is typically insensitive to reduced miRNA activity
- Extensively paired sites can reduce derepression thresholds by triggering miRNA decay
- Weak sites can contribute to target-site competition without imparting repression
- Closely spaced sites of the same or different miRNAs cooperatively sequester miRNAs

Accession Numbers

GSE76288

Denzler et al., 2016, Molecular Cell *64*, 565–579
November 3, 2016
http://dx.doi.org/10.1016/j.molcel.2016.09.027

Molecular Cell
Article

Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression

Rémy Denzler,[1] Sean E. McGeary,[2,3,4] Alexandra C. Title,[1] Vikram Agarwal,[2,3,4,5] David P. Bartel,[2,3,4,*] and Markus Stoffel[1,6,*]
[1]Institute of Molecular Health Sciences, Swiss Federal Institute of Technology in Zurich (ETH Zurich), Otto-Stern-Weg 7, 8093 Zürich, Switzerland
[2]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[3]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[4]Department of Biology
[5]Computational and Systems Biology Program
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Lead Contact
*Correspondence: dbartel@wi.mit.edu (D.P.B.), stoffel@biol.ethz.ch (M.S.)
http://dx.doi.org/10.1016/j.molcel.2016.09.027

SUMMARY

Expression changes of competing endogenous RNAs (ceRNAs) have been proposed to influence microRNA (miRNA) activity and thereby regulate other transcripts containing miRNA-binding sites. Here, we find that although miRNA levels define the extent of repression, they have little effect on the magnitude of the ceRNA expression change required to observe derepression. Canonical 6-nt sites, which typically mediate modest repression, can nonetheless compete for miRNA binding, with potency ~20% of that observed for canonical 8-nt sites. In aggregate, low-affinity/background sites also contribute to competition. Sites with extensive additional complementarity can appear as more potent, but only because they induce miRNA degradation. Cooperative binding of proximal sites for the same or different miRNAs does increase potency. These results provide quantitative insights into the stoichiometric relationship between miRNAs and target abundance, target-site spacing, and affinity requirements for ceRNA-mediated gene regulation, and the unusual circumstances in which ceRNA-mediated gene regulation might be observed.

INTRODUCTION

MicroRNA (miRNA) levels have long been known to influence the magnitude of target-gene repression (Bartel, 2009). More recent studies point out that the number of predicted binding sites present in the transcriptome also affects the activity of miRNAs (Arvey et al., 2010; Garcia et al., 2011). Consistent with this concept, strong overexpression of natural or artificial RNAs that contain miRNA sites can titrate miRNAs away from natural targets, thereby reducing the repression of these transcripts (Ebert et al., 2007; Franco-Zorrilla et al., 2007; Mukherji et al., 2011; Hansen et al., 2013; Memczak et al., 2013). These observations are extended by the notion that a site-containing transcript found naturally within cells can act as competing endogenous RNA (ceRNA) and regulate other site-containing transcripts by increasing or decreasing the miRNA activity (Poliseno et al., 2010; Cesana et al., 2011; Salmena et al., 2011; Karreth et al., 2015).

The ceRNA hypothesis remains controversial due to the lack of a plausible explanation for how modulating the expression of a single endogenous gene could perceptibly influence miRNA activity across all of its target sites. Two recent studies have empirically assessed the ceRNA hypothesis by quantifying the number of miRNA response elements (MREs) that must be added to detect ceRNA-mediated gene regulation (Bosson et al., 2014; Denzler et al., 2014). Both studies agree that determining the number of transcriptomic miRNA-binding sites is crucial for evaluating the potential for ceRNA regulation and that miRNA-binding sites are generally higher than the number of miRNA molecules. However, they differ in two aspects: (1) the experimental approaches used to determine the number of "effective" transcriptomic miRNA-binding sites and (2) the impact miRNA concentrations have on the number of binding sites that must be added to detect target gene derepression (derepression threshold [DRT]).

The discrepancies between these studies lead to different conclusions with respect to the likelihood of observing ceRNA effects in natural settings. The first study concluded that changes in ceRNAs must approach a miRNA's target abundance before they can exert a detectable effect on gene regulation (Denzler et al., 2014). Furthermore, because target abundance for a typical miRNA is very high, regulation of gene expression by ceRNAs is unlikely to occur in differentiated cells under physiological settings or most disease settings (Denzler et al., 2014). In addition, the study shows that the DRT remains constant when miRNA activity is reduced. A subsequent review presents a mathematical model that assesses binding-site occupancy and competition at different assumed target abundances (Jens and

Rajewsky, 2015). This in silico model predicts that only global and collective changes in binding sites can produce an effect on target abundance large enough to detectably derepress target genes, which concurs with the results and conclusions of Denzler et al. (2014).

The second study presents a "hierarchical affinity model," in which the miRNA abundance is proposed to determine the respective susceptibility to ceRNA-mediated regulation (Bosson et al., 2014). In this model, the suggestion is that, as miRNA concentration increases and Ago-miRNA complexes spread to weaker and weaker sites (with affinity inferred from the site hierarchy of 8-nt > 7-nt > 6-nt site), the effective target-site abundance grows too large for physiological ranges of ceRNA expression to influence repression. By this reasoning, physiological ceRNA changes can nevertheless influence repression by a more modestly expressed miRNA, with its correspondingly lower effective target-site abundance. Moreover, the use of high-throughput cross-linking to detect targets leads to lower target-abundance estimates, which further increases the plausibility of ceRNA regulation (Bosson et al., 2014). However, experimental support for the proposed influence of miRNA concentration is correlative and lacks direct experimental evidence, such as manipulation of miRNA activity and measurement of resulting DRT changes.

Denzler et al. (2014) propose that sites of all different affinities contribute to the effective target abundance, regardless of the miRNA concentration. Here, we call the model of Denzler et al. (2014) the "mixed-affinity model" to distinguish it from the hierarchical affinity model. The mixed-affinity model recognizes that a high-affinity site will contribute more to effective target-site abundance than a low-affinity site (Denzler et al., 2014). However, in aggregate, low-affinity sites, because of their high numbers within the transcriptome, still make a substantial contribution to the effective target-site abundance for each miRNA—even for more modestly expressed miRNAs.

Other studies suggest that the ceRNA crosstalk of two transcripts is stronger and more specific when they share a large number of sites to different miRNA seed families. This hypothesis emerged from observations in cancer models, in which the expression of a particular oncogene correlates with its pseudogene, and both transcripts share a high sequence homology in their 3′ UTRs and are reported to co-regulate each other through a ceRNA mechanism (Poliseno et al., 2010; Karreth et al., 2015). Even if transcripts containing multiple sites can exert an additive effect of independently acting binding sites, sites for each miRNA family would still have to individually reach the high thresholds necessary to observe target-gene derepression. Therefore the simple presence of multiple binding sites alone would not be expected to be sufficient to increase the likelihood of a ceRNA effect, unless the sites acted through a cooperative mechanism. Although the effect of cooperativity has been studied in the context of target-gene repression (Doench et al., 2003; Grimson et al., 2007; Saetrom et al., 2007; Broderick et al., 2011), it is unclear whether closely spaced miRNA-binding sites can sequester miRNA in a non-independent manner and hence increase the prospects of a ceRNA effect.

In this study, we examine the impact that miRNA levels have on the DRT and thereby address a key difference between the hierarchical affinity and mixed-affinity models. We then analyze the influence of target-site complementarity on ceRNA-mediated gene regulation and examine the extent to which closely spaced miRNA-binding sites can cooperatively influence the potency of target-gene derepression. Finally, we develop a mathematical model, which incorporates both the mixed-affinity binding and the repressive activities of miRNAs to recapitulate our results.

RESULTS

miR-294 Is Susceptible to Competition Despite High Expression Levels

A powerful tool for studying competition among MREs is a single-cell reporter assay that transcribes *mCherry* mRNA (with or without MREs in its 3′ UTR) and enhanced yellow fluorescent protein (*eYFP*) mRNA as an internal measure of reporter transcription (Mukherji et al., 2011; Bosson et al., 2014). Using analytical flow cytometry, *mCherry* reporter readout can be assessed over a broad range of added MREs. At high expression levels, MREs can compete with each other for miRNA binding, thereby causing derepression. Using this assay in embryonic stem cells (ESCs), some miRNAs need fewer competing MREs to mediate reporter derepression and are therefore more susceptible to ceRNAs than other miRNAs (Bosson et al., 2014).

To explore these different susceptibilities, we created reporter constructs for six highly expressed ESC miRNAs (miR-294, -293, -92, -16, -26, and -292-5p) (Bosson et al., 2014), containing zero (0s), or three (3s) 8-nt miRNA sites in the 3′ UTR of *mCherry* (Figure 1A). For the miRNA families miR-294, -293, and -92, reporters containing a single (1s) miRNA-binding site were also created. Sites for miR-294, -293, -92, and -16 (Figures 1B and 1C), but not those for miR-26 and -292-5p (data not shown) caused detectable miRNA-mediated repression of mCherry. The extent of repression of reporters for miR-294, -293, and -92 resembled that observed previously, as did the derepression of mCherry constructs harboring sites for miR-293 or miR-92 (Bosson et al., 2014). However, the 3s reporter construct for miR-294, a miRNA reported to be insensitive to competitor perturbations (Bosson et al., 2014), and the reporter for miR-16 were derepressed when eYFP fluorescence exceeded 2.2×10^4 or 2.8×10^4, respectively (Figure 1B). The ability to observe derepression of the miR-294 reporter presumably resulted from improvements to the equipment and protocol that enabled more precise measurements, as indicated by the improved SEM values, although differences between the ESCs might have also played a role. These results showing derepression of the mCherry reporter at similar competitor levels for both miR-294 and miR-16, two miRNAs present at very different levels in ESCs, and with very different miRNA:target ratios estimated by Bosson et al. (2014), support the mixed-affinity model.

Derepression of Target mRNAs Occurs at a High Threshold of Added Target Sites

The competition among MREs for miRNA binding is expected to occur not only between the added MREs within the *mCherry* mRNA but also between the added MREs and those of the

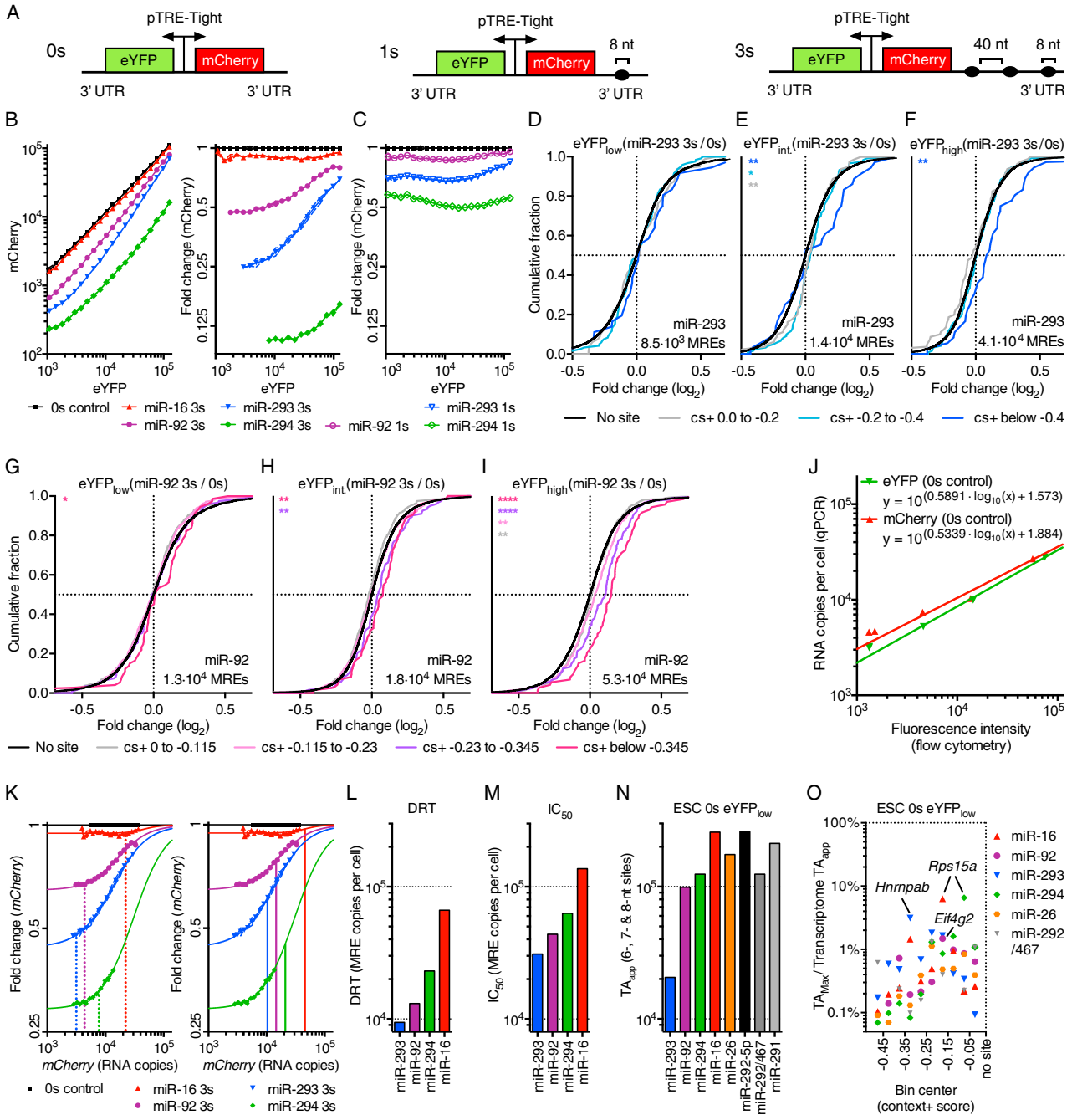


Figure 1. Derepression of Target mRNAs Occurs at a High Threshold of Added Target Sites

(A) Dual-color fluorescent reporter constructs containing zero (0s), one (1s), or three (3s) 8-nt miRNA site(s) in the 3′ UTR of *mCherry*.

(B and C) ESCs transfected with either a 3s (B), 1s (C), or 0s reporter construct (n = 3) with miRNA-binding sites for miR-294, -293, -92, or -16. Mean mCherry fluorescence (B, left), and mCherry fluorescence normalized to the 0s control (B, right and C) across 20 bins of eYFP.

(D–I) RNA-seq results (n = 2) of sorted ESCs shown in Figure S1A. ESCs were transfected with a 3s reporter for miR-293 (D–F) or miR-92 (G–I), or a 0s control, and gated for cells with low ($eYFP_{low}$) (D and G), intermediate ($eYFP_{int.}$) (E and H), or high eYFP ($eYFP_{high}$) (F and I) expression. Cumulative distribution function (CDF) of mRNA changes for predicted target genes with the indicated context+ score (cs+) bins (color) or for genes with no miRNA site (black). *mCherry* MREs per cell evaluated by qPCR are shown on each graph. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, one-sided Kolmogorov-Smirnov (K-S) test. Also see Figures S1B and S1C.

(J) Relationship between reporter protein fluorescence measured by flow cytometry and RNA copies per cell evaluated by qPCR of ESCs transfected with the 0s reporter and sorted into four different bins of eYFP-expressing cells. Line represents non-linear regression of data points; respective equations are shown.

(legend continued on next page)

endogenous targets. To examine the effect on endogenous targets, ESCs transfected with the 0s or a 3s reporter for either miR-293 or miR-92 were sorted into three bins based on their eYFP expression (Figure S1A, available online). RNA sequencing (RNA-seq) of each bin revealed the number of MREs added per cell as well as differences in endogenous mRNA levels for cells with the 3s reporter compared to those with the 0s reporter. Endogenous mRNAs with predicted MREs were grouped based on the strength of their predicted response to the miRNA, as scored by the context+ model of TargetScan 6.2 (Garcia et al., 2011). For the middle, but not the lower, bin (1.4×10^4 and 0.85×10^4 added miR-293 MREs per cell, respectively), endogenous miR-293 targets were derepressed, as indicated by the significant shift in the distribution of mRNA fold-change values of the top predicted miR-293 targets (Figures 1D–1F and S1D–S1F; Table S1). Likewise, convincing miR-92 target derepression was not observed until exceeding 1.3×10^4 added miR-92 MREs (Figures 1G–1I and S1G–S1I).

Comparison of mCherry and eYFP fluorescence with the corresponding transcript copy numbers, as measured by qRT-PCR (qPCR), revealed that fluorescence and mRNA abundance were highly correlated, although the relationship was not one-to-one (Figure 1J). Because protein fluorescence intensity is an indirect readout that is not directly relevant to the competition that occurs on the level of mRNA and miRNA, we transformed the fluorescence values measured by flow cytometry in Figure 1B to transcript copies per cell by employing the standard curves of Figure 1J (Figures 1K and S1J). Strikingly, the DRT observed for miR-293 and miR-92 reporters (0.9×10^4 and 1.3×10^4 sites per cell, respectively; Figures 1K and 1L) resembled those observed by RNA-seq for endogenous targets, thereby validating the reporter output (after transforming fluorescence to transcript copy number) for endogenous target derepression.

We next calculated the number of MREs that must be added per cell to observe half-maximal derepression (termed half-maximal inhibitory concentration, or IC_{50}) of the different reporter constructs (Figures 1K and 1M). The number of miRNA molecules per ESC is reported to be 5.7×10^4 for miR-294, 2.6×10^3 for miR-293, 1.7×10^3 for miR-92, and 1.8×10^3 for miR-16 (Bosson et al., 2014), which was consistent with the relative levels of these miRNAs in our ESCs, as determined by small-RNA-seq (Figure S1L; Table S2). Thus, as observed for miR-122 in hepatocytes (Denzler et al., 2014), the IC_{50} values exceeded the number of miRNA molecules per ESC. In such a regime, the IC_{50} provides an empirical measure of the effective endogenous target-site abundance, as half-maximal derepression should be achieved when the competing sites reach an effective concentration matching that of the endogenous sites (Denzler et al., 2014).

In hepatocytes, the miR-122 IC_{50} (4.5×10^5 sites per cell) happens to correspond to the sum of all 3′ UTR 6-, 7-, and 8-nt sites of the transcriptome, leading to the idea that this sum, defined as the TA_{app}, can provide an estimate of the effective target-site abundance for other miRNAs (Denzler et al., 2014). To test this idea, we examined the correspondence between the newly determined IC_{50} values and the TA_{app} values for the ESC transcriptome. When comparing RNA-seq data with absolute copy numbers of *mCherry*, *eYFP*, and three differently expressed genes, a linear association was observed (Figure S1K), which provided a standard curve to transform RNA-seq data to absolute mRNA copies per cell, enabling TA_{app} values for eight active ESC miRNAs to be determined (Figure 1N). For all four miRNAs with IC_{50} values, the TA_{app} approached the IC_{50}, ranging from ~2-fold above the IC_{50} (miR-16, -92, and -294), to 1.5-fold below the IC_{50} (miR-293). Because TA values estimated from cross-linking (Bosson et al., 2014) strongly correlated with TA_{app} values (Figures S1M and S1N), but were ~7-fold lower, the cross-linking immunoprecipitation (CLIP)-estimated TA values were not more informative for the purposes of estimating the effective target-site abundance. We conclude that summing of 3′ UTR 6-, 7-, and 8-nt sites in the transcriptome provides a reasonable approximation of effective abundance of endogenous target sites.

The DRTs ranged between 12% (miR-92) and 30% (miR-293) of TA_{app}. Importantly, no endogenous transcript contributed such a large percentage to transcriptome TA_{app} of the ESC miRNAs examined. The largest contributor was ribosomal protein S15A (*Rps15a*) mRNA, which contributed 6.5% of the TA_{app} for both miR-294 and miR-16 (Figure 1O). Thus in ESCs, as in hepatocytes (Denzler et al., 2014), ceRNA-regulated gene expression through upregulation or downregulation of a single transcript is unlikely. Similar results have been reported in HEK293 cells (Yuan et al., 2015).

Derepression Threshold Values Are Insensitive to Changes in miRNA Activity

A key difference between the mixed-affinity and the hierarchical affinity models is the impact that miRNA levels have on the threshold required to detect derepression of target genes (Bosson et al., 2014; Denzler et al., 2014). To investigate this issue, we examined the influence that reduced miRNA activity has on the DRT in the single-cell assay. ESCs were transfected with either 0s or 3s miR-293 reporters, in addition to different concentrations of Antagomir-293 (Ant-293). Reduction of mCherry repression correlated with increasing Ant-293 concentrations, confirming that miR-293 activity was reduced in Antagomir-treated ESCs (Figure 2A). As observed for miR-122 in hepatocytes (Denzler et al., 2014), the DRTs and IC_{50} values did not decrease as miR-293 activity was reduced in ESCs (Figures 2B and S2A–S2C).

(K) Protein fluorescence values shown in (B) transformed to RNA copies per cell using the equations shown in (J). Vertical lines represent the DRT (dotted lines) or IC_{50} (solid lines).

(L and M) Bar plot of DRT (L) and IC_{50} (M) shown in (K).

(N) Transcriptome TA_{app} of ESCs transfected with the 0s reporter and sorted for low eYFP-expressing cells (ESC 0 s $eYFP_{low}$).

(O) Fractional contribution of the largest potential contributors to transcriptome TA_{app} of ESC 0s $eYFP_{low}$. Potential contributors were binned by their context+ score, and the top potential contributors are plotted within each bin.

Data represent mean ± SEM for (B), (C), and (K).

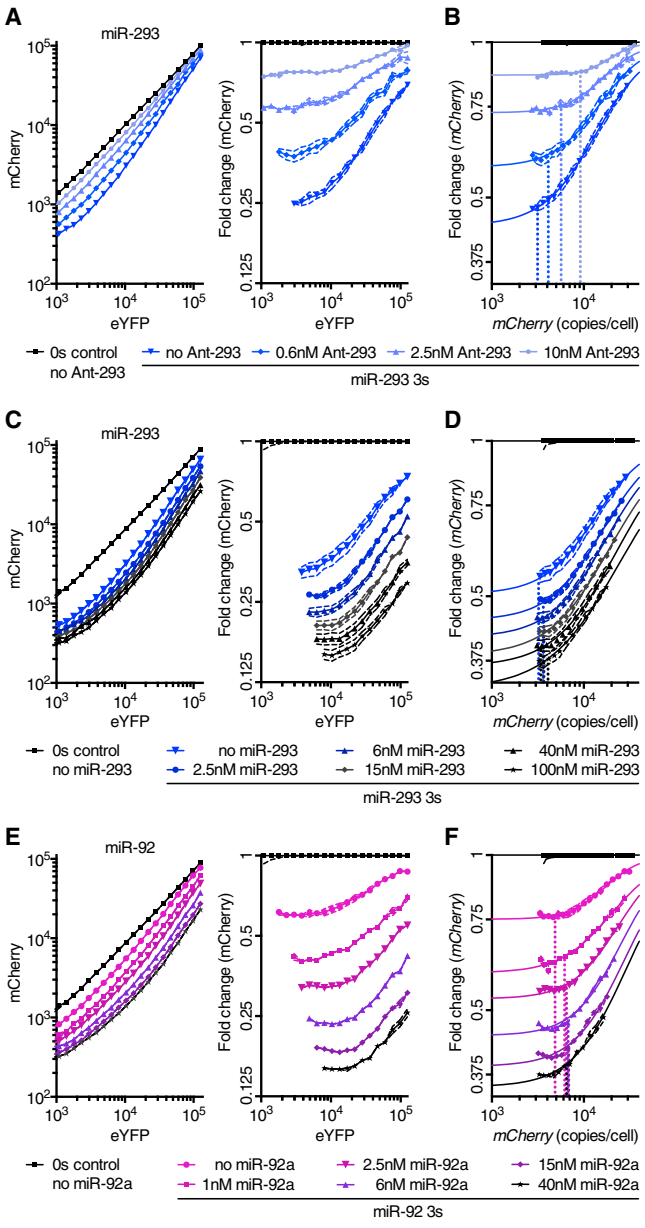


Figure 2. Derepression Threshold Values Are Insensitive to Changes in miRNA Activity

(A–F) ESCs co-transfected with a 3s reporter for miR-293 (A–D), miR-92 (E and F), or respective 0s reporter control, and different concentrations of Ant-293 (n = 3) (A and B), miR-293 (n = 6) (C and D), or miR-92 (n = 6) (E and F).

(A, C, and E) Mean mCherry fluorescence (left), and mCherry fluorescence normalized to the 0s control (right) across 20 bins of eYFP.

(B, D, and F) Protein fluorescence values shown in (A), (C), and (E) were transformed to RNA copies per cell using the equations shown in Figure 1J. Vertical, dotted lines denote the DRT.

Data represent mean ± SEM for all panels.

We next increased miRNA activity and examined the effect on the DRT. ESCs were transfected with the dual-fluorescent reporter and different concentrations of miRNA duplex. When quantified with respect to eYFP fluorescence or *eYFP* mRNA copies, we detected an increase in the DRT as more miRNA was transfected (Figures 2C, 2E, S2D, S2E, S2G, and S2H). However, eYFP, unlike mCherry, is not a good measure for MRE induction as it is not repressed by the miRNA and hence does not inform how many MREs are actually expressed in a cell. Indeed, when quantified with respect to *mCherry* transcript abundance, the DRT of competing transcripts remained constant as more miRNA was transfected (Figures 2D, 2F, S2F, and S2I). These results monitoring DRTs after decreasing or increasing miRNA activities supported the mixed-affinity model, in which less abundant miRNAs should be no more susceptible to ceRNA effects than are more abundant miRNAs (Denzler et al., 2014).

Extensively Paired Sites Are More Potent Than 8-nt Sites and Trigger miRNA Decay

We investigated whether the DRT was also insensitive to increased miRNA levels in primary hepatocytes. A 4-fold increase in miR-122, attained by infecting hepatocytes with a recombinant adenovirus expressing the miR-122 precursor (Ad-miR-122), resulted in decreased levels of endogenous miR-122 target mRNAs (Figures S3A–S3C). To manipulate miR-122 MREs and measure the subsequent effects on miR-122 target genes, we increased the levels of the miR-122 target *AldolaseA* (*AldoA*) mRNA using an adenovirus (Ad-AldoA) that carried either a mutated site (Mut), one (1s), or three sites (3s) to miR-122 (Figure 3A). Hepatocytes were infected at different multiplicities of infection (MOIs), at either basal or elevated miR-122 levels (Figures 3B and S3D–S3G). At endogenous miR-122 levels, we began to observe miR-122 target derepression when more than 2.1×10^5 miR-122 MREs were introduced (Figure 3C). The DRT did not increase when endogenous miR-122 levels were raised 4-fold (Figure 3C), which is in agreement with our observations in ESCs. Of note, the higher DRT observed in hepatocytes compared to ESCs is expected based on the larger cytoplasm and number of mRNAs per cell in hepatocytes.

Our finding that derepression occurred at only high thresholds of added target sites seemed to disagree with a study in HeLa cells that used "bulged" binding sites with near-perfect complementarity (Mukherji et al., 2011). We sought to test the possibility that sites with perfect complementarity to the miRNA 3′ region might yield different results because they mediate miRNA degradation. Hepatocytes were infected with Ad-AldoA containing either a mutated or a bulged (bu4) binding site (Figure 3A). Interestingly, derepression was already observed when exceeding only 5×10^4 bulged miR-122 MREs per cell (Figures 3D, 3E, S3H, and S3I), confirming that bulged sites are more efficient than 8-nt sites in influencing miRNA activity. The efficiency of target-mRNA derepression mediated by bulged sites correlated well with a decrease of miR-122, but not miR-16, levels (Figures 3F and S3J), suggesting that derepression was induced by enhanced miRNA degradation rather than direct competition between miRNA-binding sites.

Target-mediated miRNA decay is associated with tailing and trimming of the miRNA (Ameres et al., 2010). Indeed, we observed reduced miR-122 signal with evidence of tailing and trimming when bulged, but not 8-nt, seed matches caused target-gene derepression (Figures 3G and S3K). These results

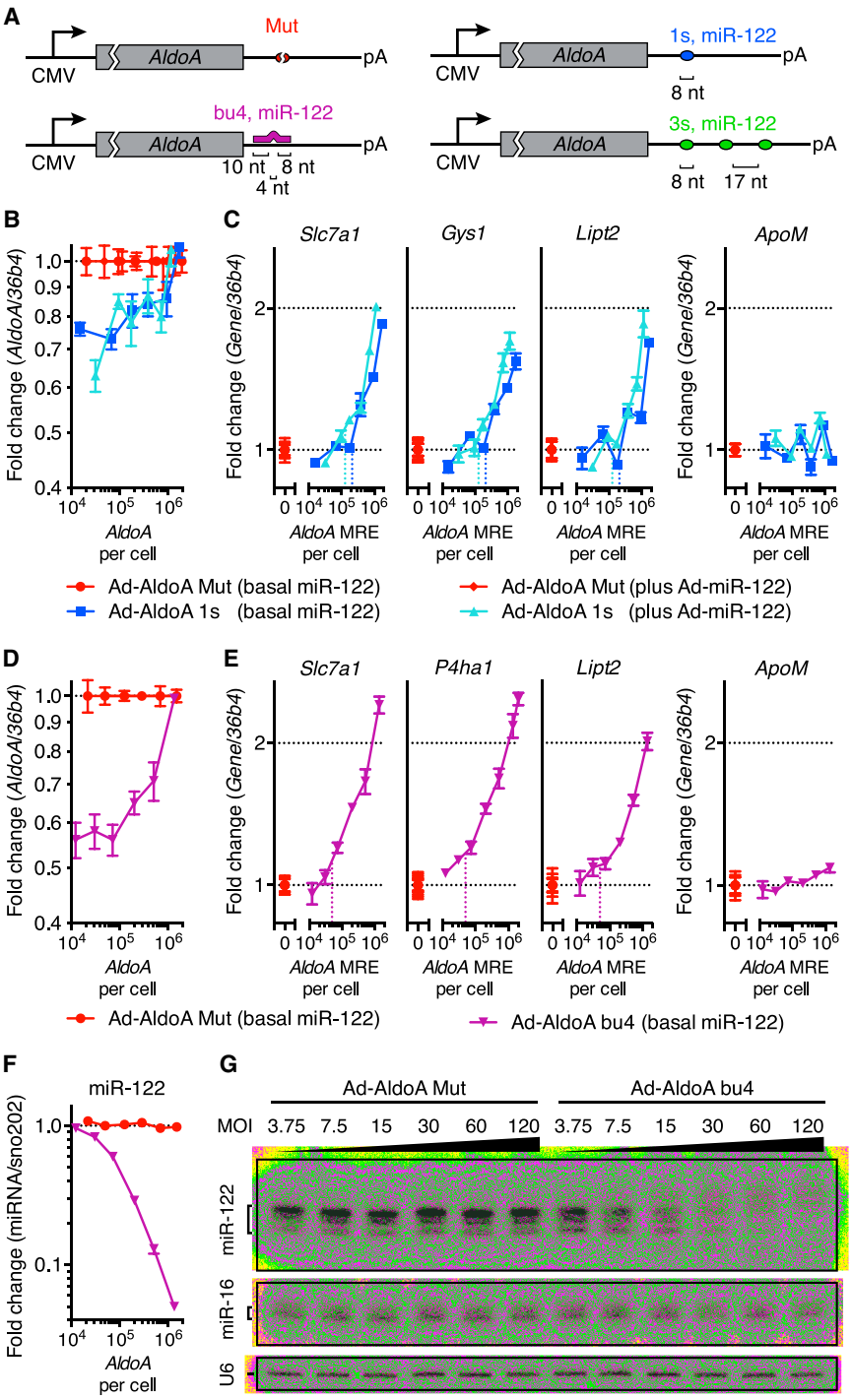


Figure 3. Extensively Paired Sites Are More Potent Than 8-nt Sites and Trigger miRNA Decay

(A) Schematic overview of the different *AldoA*-expressing adenovirus constructs.

(B–G) Primary hepatocytes (n = 4) infected with different MOIs of Ad-AldoA 1s (B and C), bu4 (D–G), or respective Ad-AldoA Mut controls at either basal miR-122 levels (B–G) or with co-infected Ad-miR-122 (B and C). Relative levels of *AldoA* (B and D), miR-122 target genes and control non-target gene (*ApoM*) (C and E), or miR-122 (F). Vertical, dotted lines denote the DRT. miRNA levels are relative to the lowest MOI of Ad-AldoA Mut at basal miR-122 levels. (G) Northern blot analysis of miR-122, miR-16, and U6 at basal miR-122 levels.

Data represent mean ± SEM for all panels. Also see Figure S7.

confirmed that bulged sites with perfect complementarity to the miR-122 3′ region reduce miRNA activity primarily through miRNA degradation rather than competition with other binding sites. Therefore, to be effective, these bulged sites need not approach the effective abundance of the miRNA target sites, but need only to be sufficiently abundant that the amount of target-mediated RNA decay substantially decreases the miRNA abundance.

miRNA Target Derepression for let-7, miR-194, and miR-192 Also Occurs at a High Threshold of Added MREs

To consider the susceptibility of other hepatocyte miRNAs to ceRNA-mediated gene regulation, we first measured absolute levels of miR-122 and six other miRNA seed families highly expressed in liver (Denzler et al., 2014). These levels ranged from 3.8×10^3 to 1.4×10^5 copies per cell (Figures 4A and S4A) and correlated well with small-RNA-seq data (Figure S4B; Table S2). We selected four families (let-7, miR-194, -192, and -101) that were not influenced by control virus expression (Figure S4C) and were expressed above 1.8×10^4 copies per cell. To study the sensitivity of these four miRNA families to competing RNA perturbations, Ad-AldoA constructs were generated in which the miR-122 site was replaced with a single 8-nt site (1s) for the respective miRNA (Figure 4B). We first infected hepatocytes with different MOIs of Ad-AldoA Mut or 1 s (let-7). Derepression of let-7 targets, which were validated by transfection of let-7f mimics (Figures S4D–S4G), was observed when >2.1×10^5 let-7 MREs were expressed per cell (Figures 4C, S4H, and S4I). This DRT was consistent with RNA-seq results (Figures 4D–4F and S4J–S4L; Table S3). In contrast, addition of up to 10^6 MREs of either miR-192, miR-194, or miR-101 through respective Ad-AldoA infections did not result in detectable derepression of validated targets (Figures S4M–S4P; data not shown), suggesting that the endogenous level of 1.8×10^5 miRNA molecules per cell did not impart sufficient repression upon which derepression could act. We therefore performed the analogous

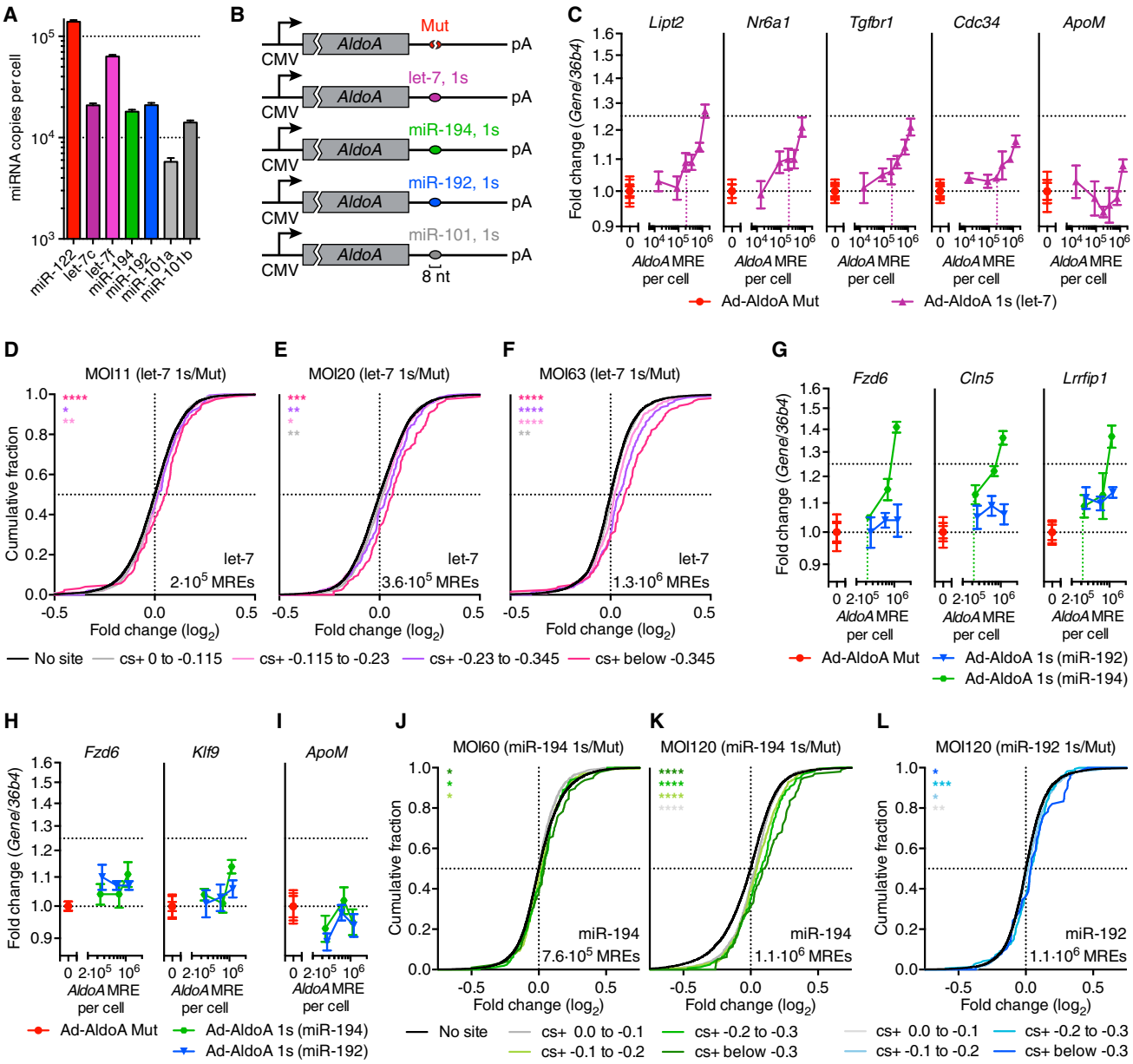


Figure 4. miRNA Target Derepression for let-7, miR-194, and miR-192 Also Occurs at a High Threshold of Added MREs

(A) Absolute copies per cell of hepatocyte miRNAs.

(B) Schematic overview of Ad-AldoA constructs harboring a mutated site (Mut), or one (1s) 8-nt binding site for let-7, miR-194, -192, or -101.

(C–F) Primary hepatocytes infected with different MOIs of Ad-AldoA Mut or 1s (let-7).

(C) Relative expression of let-7 target genes and control non-target gene (*ApoM*).

(D–F) CDF of RNA-seq data (n = 2) showing mRNA changes for predicted target genes of let-7 with the indicated cs+ bins (color) or for transcripts with no miRNA site (black).

(G–L) Hepatocytes infected with different MOIs of Ad-AldoA Mut, 1s (miR-192), or 1s (miR-194), in addition to MOI 15 Ad-miR-192/194. Relative levels of miR-194 (G) or miR-192 (H) target genes, and control non-target gene (*ApoM*) (I). CDF of RNA-seq data (n = 2) showing mRNA changes for predicted target genes of miR-194 (J and K) or miR-192 (L) with the indicated cs+ bins (color) or for genes with no miRNA site (black).

AldoA MREs per cell evaluated by qPCR are shown on each graph. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, one-sided K-S test. Vertical, dotted lines denote the DRT.

Data represent mean ± SEM (n = 4) for all panels.

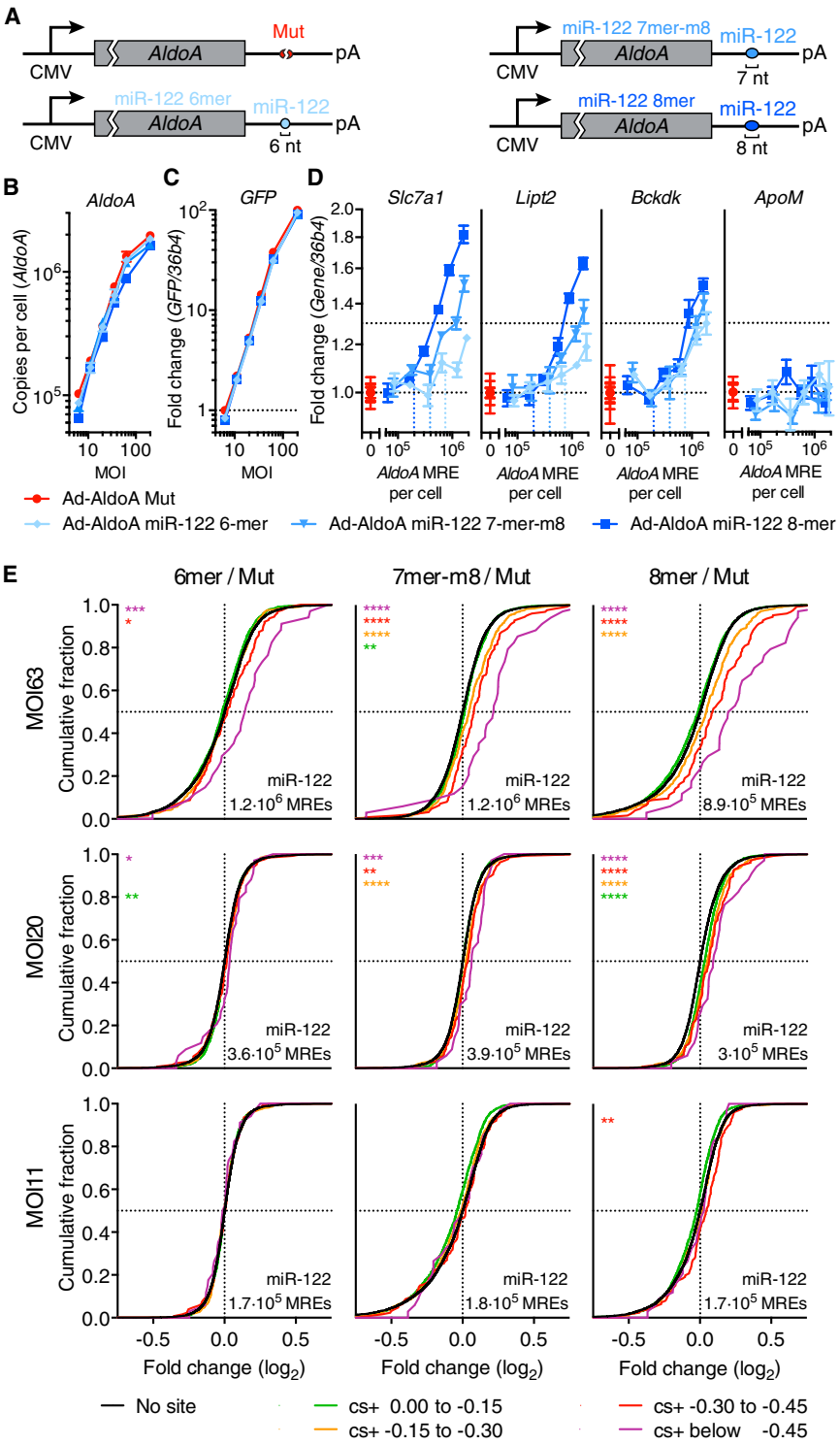


Figure 5. The 6-, 7-, and 8-nt Sites Contribute Comparably to Target Abundance of miR-122

(A) Schematic overview of Ad-AldoA constructs used in this figure.

(B–E) Primary hepatocytes infected with different MOIs of Ad-AldoA miR-122 8-mer, miR-122 7-mer-m8, miR-122 6-mer, or Mut. Absolute copy numbers per cell of *AldoA* (B), relative gene expression of *GFP* (C), and of miR-122 target genes or control non-target gene (*ApoM*) (D). (E) CDF of RNA-seq data (n = 2) showing mRNA changes for predicted target genes of miR-122 with the indicated cs+ bins (color) or for genes with no respective miRNA site (black). *AldoA* MREs per cell evaluated by qPCR are shown on each graph. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, one-sided K-S test. See also Figures S5C and S5D.

Vertical, dotted lines denote the DRT. Data represent mean ± SEM (n = 4) for all panels.

experiment under conditions of elevated miR-192 and miR-194 levels using recombinant adenovirus expression (Ad-miR-192/194) (Figure S4M). Increasing miR-194 by 4.5-fold increased repression to a level at which target derepression could be observed, with a DRT of >3.2 × 10^5 added miR-194 MREs per cell (Figures 4G, 4I, S4Q, and S4R). RNA-seq analysis confirmed a similar DRT for predicted miR-194 targets (Figures 4J, 4K, S4S, and S4T). Because repression of predicted targets was not readily observed when increasing miR-192 levels by 3.7-fold (Figure S4O), derepression was also difficult to measure (Figures 4H, 4I, S4Q, and S4R), although some signal for derepression was detected when 1.1 × 10^6 miR-192 MREs were added (Figures 4L and S4U). Together, these results indicated that derepression of let-7, miR-192, and miR-194 targets in hepatocytes occurred at similar or higher DRTs than previously observed for miR-122 targets.

The 6-, 7-, and 8-nt Sites Contribute Comparably to Target Abundance

The different levels of repression efficacy and preferential conservation observed for 6-, 7-, and 8-nt site types (Bartel, 2009) raised the question as to the extent to which these site types differ in their efficacy as competitors. Accordingly, we infected hepatocytes with different MOIs of Ad-AldoA constructs harboring either a mutated or one 6-, 7-, or 8-nt site to miR-122 (Figure 5A). Derepression of miR-122 targets was observed when adding each of the three site types, with a clear relationship between competitor site type and DRT (Figures 5B–5D). This relationship, in which DRT increased as site size decreased, was also observed when extending our analysis to the transcriptome (Figures 5E and S5A). For example, the derepression observed for the 6-nt site at MOI 63 was between that

observed for the 8-nt site at MOI 11 and 20, suggesting that as a competitor it was about 20% as effective as the 8-nt site. With respect to the 7-nt site, derepression at MOI 63 exceeded that observed for the 8-nt site at MOI 20, and derepression at MOI 20 surpassed that measured for the 8-nt site at MOI 11, suggesting that as a competitor the 7-nt site was about 50% as effective as the 8-nt site. Employing these factors to calculate a weighted TA_{app} only decreased the TA_{app}, without affecting the relative ranking of the respective miRNA TA_{app} (Figure S5B). These results indicate that, in aggregate, 7-nt sites, which are 3- to 8-fold more abundant than 8-nt sites, contribute more to effective target-site abundance than do 8-nt sites, and that 6-nt sites contribute more to effective target-site abundance than might have been expected from their marginal efficacy in target repression.

Derepression Is Enhanced When Mediated by Closely Spaced MREs

Although the cooperative effect of closely spaced miRNA-binding sites has been studied in the context of mRNA repression (Doench et al., 2003; Grimson et al., 2007; Saetrom et al., 2007; Broderick et al., 2011), the role of cooperatively spaced miRNA-binding sites has not been investigated in the setting of site competition. We therefore analyzed whether closely spaced miRNA-binding sites can cooperatively sequester miRNA molecules and hence reduce the number of sites required for derepression.

Cooperatively acting MREs within endogenous 3′ UTRs tend to be between 8 and ~60 nt apart (Grimson et al., 2007; Saetrom et al., 2007). We thus generated Ad-AldoA constructs harboring one 8-nt site for miR-122 and one for let-7, separated by 58 nt (Ad-AldoA 2x +58nt), or respective single-site controls (Figure 6A), and infected hepatocytes at different MOIs. Interestingly, predicted let-7 targets that lacked miR-122 sites showed stronger derepression when let-7 MREs were added through constructs harboring a nearby miR-122 site (Figures 6B–6D, S6A, and S6B). Analogous results were obtained for the derepression of miR-122 targets by miR-122 MREs that had an adjacent let-7 site, showing that competition for binding to one miRNA family can be influenced by a nearby site of a different family. To achieve the same level of derepression conferred by isolated sites, the sites with nearby cooperative sites required only 20%–50% as many molecules per cell (Figure 6B).

To study the influence that the spacing of the miR-122 and let-7 sites has on the ability to cause cooperative competition, we infected hepatocytes with various MOIs of differently spaced Ad-AldoA 2x constructs (Figure 6A). Although the cooperative effect of Ad-AldoA-2x-mediated gene regulation persisted independently of whether the let-7–binding site was 58 nt upstream or downstream of the miR-122 site—indicating that a specific intervening sequence or structure was not required—no cooperative effect was observed when the two sites were 255 or 997 nt apart (Figures 6E–6G, S6C, and S6D). When changing the 8-nt miR-122 site on Ad-AldoA 2x +58nt to a 7- or 6-nt site (Figure 6A), a strong relationship was observed between site type and the magnitude of the cooperative effect (Figures 6H, 6I, and S6E–S6G). Moreover, at the transcriptome level, predicted let-7 targets that lacked predicted miR-122 sites were significantly more derepressed if the competing let-7 site had an adjacent miR-122 site (Figures 6J and S6H), thereby confirming that closely spaced binding sites of co-expressed miRNAs can boost the efficacy of competing sites.

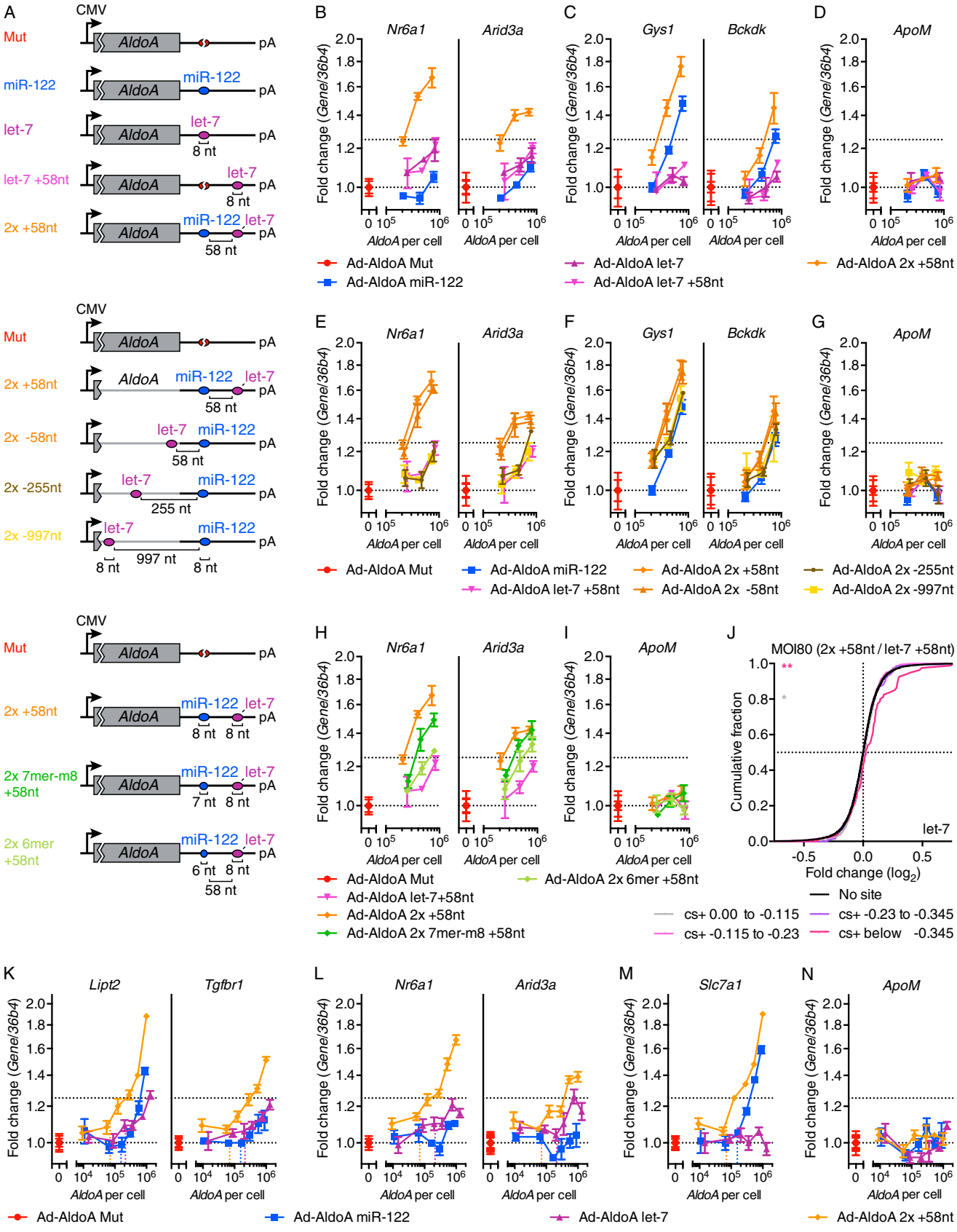
We then investigated whether the DRT was lower in conditions in which cooperativity was present by infecting hepatocytes with different MOIs using Ad-AldoA miR-122, let-7, and 2x +58nt. Derepression of miR-122 and let-7 target genes was detected when 7×10^4 *AldoA* copies of the 2x +58nt construct were exceeded (Figures 6K–6N, S6I, and S6J). Depending on whether binding sites of either miR-122 or let-7 alone, or both together are included in the cooperative DRT, the 7×10^4 *AldoA* copies would correspond to a DRT of either 7 or 14×10^4 MREs, which is either 3.1- or 1.5-fold lower (let-7), or 2.2- or 1.1-fold lower (miR-122), respectively, than the previously determined DRTs. Regardless of the DRT interpretation, these results indicate that the cooperative action of sites can detectibly boost the prospects of ceRNA-mediated gene derepression.

Mathematical Framework for the Mixed-Affinity Model

Mathematical simulations of miRNA-target interactions have been used to evaluate the potential effects of competing MREs (Mukherji et al., 2011; Ala et al., 2013; Bosson et al., 2014; Jens and Rajewsky, 2015; Schmiedel et al., 2015). However, these simulations either model only the extent to which different site types are occupied by a miRNA without modeling repressive effects that are needed for comparison to experimental results (Bosson et al., 2014; Jens and Rajewsky, 2015), or they model the repression of mRNA from one or two genes without modeling competition of sites from other expressed transcripts (Mukherji et al., 2011; Ala et al., 2013; Schmiedel et al., 2015). These latter simulations also omit the bound form of the mRNA from its simulated abundance. Most importantly, previous simulations also ignore the influence of the large number of low-affinity, non-canonical/background sites.

We therefore built a mathematical framework that incorporates site-type occupancy, mRNA destabilization, and the range of binding-site affinities intrinsic to the mixed-affinity model. This framework was used to predict the influence of both target-site abundance and miRNA level on target derepression. As with previous simulations (Mukherji et al., 2011; Bosson et al., 2014; Jens and Rajewsky, 2015), we assumed that (1) molecular species are well mixed within the cytosol and their concentrations are not influenced by cell growth and division; (2) each mRNA and miRNA is produced at a constant rate, and unbound mRNAs and miRNAs undergo constant first-order decay; (3) upon association with a miRNA, the mRNA degradation rate increases, regardless of the site type; and (4) miRNA binding is reversible, and upon miRNA dissociation the mRNA degradation rate reverts to its original value. We also assumed that the Michaelis constant (K_M) describing mRNA degradation with respect to the miRNA-mRNA complex is well approximated by the complex dissociation constant (K_D), and that both bound and unbound mRNA are translated.

We first simulated the results of adding the 1s reporter for miR-293 to ESCs, as done in Figure 1C, setting levels of miR-293 and its canonical 3′ UTR sites to those measured by sequencing. Binding affinities of 6-, 7-, and 8-nt sites were



(legend on next page)

modeled with distributions centering on their measured affinities, and a distribution of low-affinity sites was added such that the simulated IC_{50} reflected the experimentally determined value of $\sim 3 \times 10^4$ copies per cell (Figure 7A). With this target-site distribution (Figure 7A, right), the simulation recapitulated other features of our results. For example, DRT values were only marginally sensitive to 10-fold changes in miRNA (Figure 7A, left), and this sensitivity seemed greater when plotted as a function of *eYFP*, the co-expressed mRNA lacking a miR-293 site (Figure 7A, middle). Moreover, the *mCherry* IC_{50} values were even less sensitive to miRNA changes (Figure 7A, left) and corresponded to the half-maximal occupancy values (Figure 7B).

Plotting the competition in terms of site occupancy (Figure 7B) allowed comparison to previous simulations that do not consider mRNA repression. Reconstructing the simulation of miR-293 binding in ESCs from Bosson et al. (2014), using their values for site affinity and abundance, showed that sensitivity to additional 8-nt sites was much greater than that observed in our experiments, as was the influence of miRNA levels on half-occupancy values (Figure 7C, left). Similar results were observed when applying the model of Jens and Rajewsky (2015), which uses the same mathematical framework as Bosson et al. (2014) but a continuous distribution of canonical site affinities (Figure 7C, right). Remarkably, after adding low-affinity sites such that the half-maximal occupancy value matched that inferred from our experimental results, both of the previous frameworks behaved indistinguishably from ours (Figures 7B and 7D). Thus, the fundamental difference between the mixed-affinity model and the other models, which enables our simulation to better match the experimental results, is the greater effective target abundance that results from consideration of many low-affinity sites.

DISCUSSION

Our results support the mixed-affinity model for miRNA site competition. In agreement with this model, we found that DRTs did not correlate with endogenous miRNA abundance and changed only modestly with experimental manipulations that increased or decreased miRNA levels. Because reducing miRNA levels does not substantially reduce the very high number of added MREs that are necessary to impart detectable derepression, changes in ceRNAs are not more likely to influence targets of miRNAs expressed at lower levels. Thus, the previous conclusion that a ceRNA effect on miR-122 targets in hepatocytes is unlikely to occur in normal physiological or disease conditions (Denzler et al., 2014) can now be more confidently extended to targets of other miRNAs in other cell types. Indeed, using two different cell types, testing several different miRNA families, and employing complementary single-transcript and high-throughput methods, we found that competing sites must approach ~10%–40% of a miRNA's TA_{app} in order to detectably influence miRNA activity. As nearly all transcripts each contribute <5% to TA_{app}, ceRNA-mediated gene regulation is very unlikely to occur under normal homeostatic conditions.

In disfavoring the hierarchical affinity model for site competition, we are not questioning the biochemical fact that some sites have more affinity than others, and thus low- and high-affinity sites exhibit differential occupancy. Indeed, although we disfavor the hierarchical affinity model with respect to site competition, it is nonetheless useful for explaining miRNA-mediated repression: when a miRNA is lowly expressed, only the highest-affinity sites are sufficiently occupied to mediate repression, but as miRNA expression increases, more and more intermediate- and low-affinity sites have occupancies sufficient to mediate repression. This model for repression is consistent with conclusions from cross-linking studies as well as those from mRNA-profiling studies showing a strong signal for derepression at 6-nt sites after loss of very highly expressed miRNAs (Giraldez et al., 2006; Bosson et al., 2014). The difference between modeling repression and modeling competition is that weak sites (including 6-nt, non-canonical, and background sites) all compete for binding even if they impart marginal or negligible repression and, importantly, this competition occurs regardless of the miRNA level. Although occupancy at any individual weak site is low, it cannot be discounted when modeling competition because weak, low-occupancy sites are in vast excess over high-affinity sites. The idea that these weak sites make a substantial contribution to effective target abundance is supported by our mathematical modeling showing that experimental results cannot be accurately simulated without considering the aggregate contribution of low-affinity sites. Also supporting this idea are single-molecule results showing that 6-nt sites and even some sites with only partial seed matches associate with the miRNA silencing complex at rates resembling those of the higher-affinity sites (Chandradoss et al., 2015; Salomon et al., 2015). Thus, even a miRNA expressed at a very low level, such as one molecule per cell, is expected to sample very many weak sites before (and after) occupying a high-affinity site.

Although miRNA levels do not affect the DRT, miRNA levels are important insomuch as they define the magnitude at which targets are initially repressed and hence the magnitude of effect that could theoretically be observed upon changes in ceRNA expression. Thus, ceRNA-regulated gene expression is

Figure 6. Derepression Is Enhanced When Mediated by Closely Spaced MREs

(A) Schematic overview of Ad-AldoA constructs used in this figure.

(B–J) Primary hepatocytes infected with different MOIs of Ad-AldoA constructs shown in (A). Relative gene expression of let-7 target genes (B, E, and H), miR-122 target genes (C and F), or control non-target gene (*ApoM*) (D, G, and I). (J) CDF of RNA-seq results (n = 2) showing mRNA changes from hepatocytes infected with MOI 80 of Ad-AldoA let-7 +58nt or 2x +58nt for predicted target genes of let-7 (with no predicted target sites for miR-122) with the indicated cs+ bins (color) or for genes with no let-7 or miR-122 miRNA sites (black). *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001, one-sided K-S test.

(K–N) Hepatocytes infected with different MOIs of Ad-AldoA Mut, miR-122, let-7, or 2x +58nt. Relative gene expression of predicted target genes for both let-7 and miR-122 (K), let-7 target genes (L), a miR-122 target gene (M), or a control non-target gene (*ApoM*) (N). Vertical, dotted lines denote the DRT.

Data represent mean ± SEM (n = 4) for all panels.

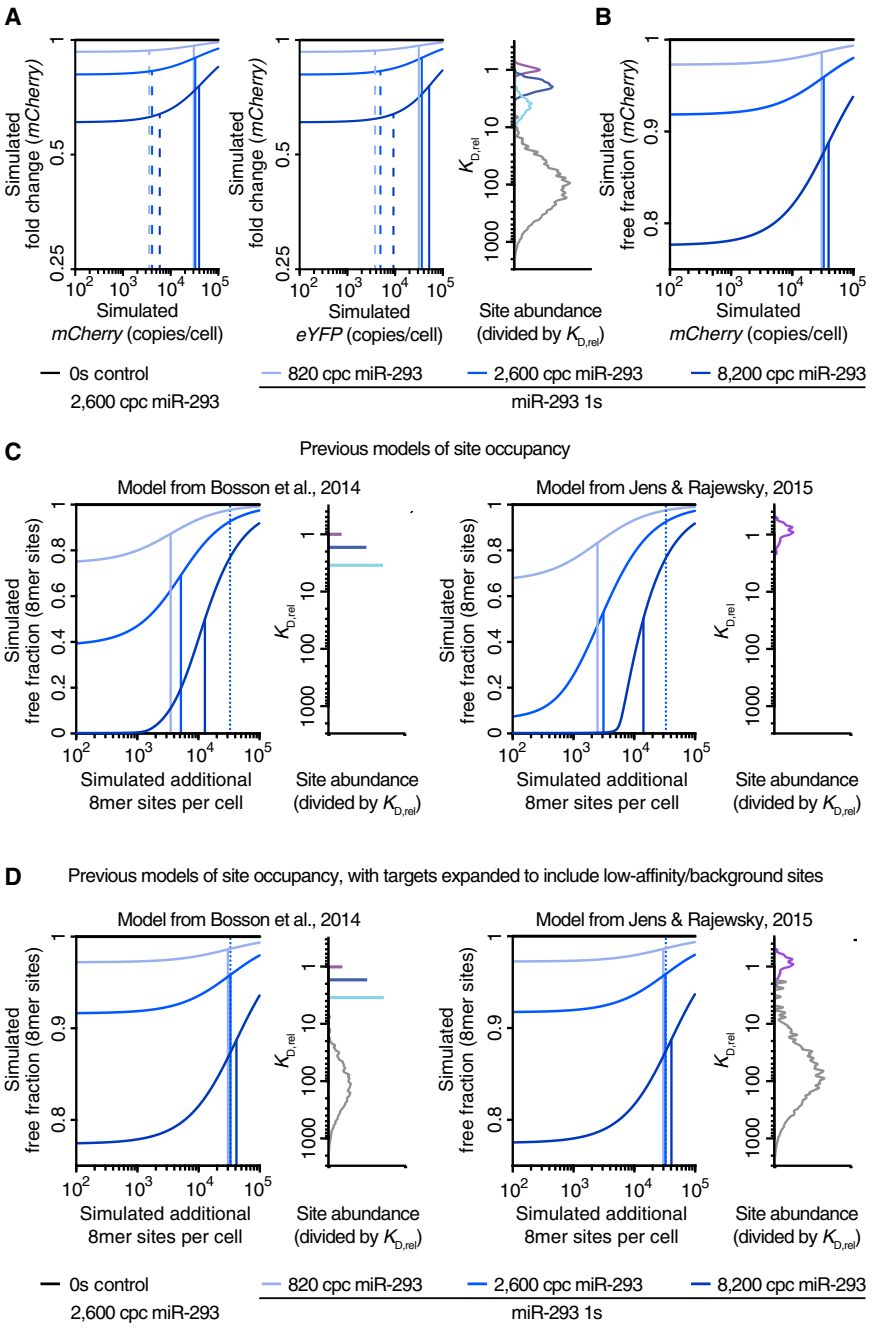


Figure 7. Mathematical Simulation of the Mixed-Affinity Model

(A) Simulated effects of changing miR-293 concentrations in ESCs on 8-nt target site repression (as performed in Figure 2), using the mathematical framework of the mixed-affinity model. Simulated *mCherry* fold-changes of the 1s reporter normalized to the 0s control are either plotted against *mCherry* (left) or *eYFP* (middle), indicating the IC_{50} (solid lines) and DRT (dashed lines), for each of the three simulated miR-293 levels (in copies per cell [cpc]). Also plotted is the binding affinity distribution of all simulated target sites (right), with the K_D of each site normalized to that of an 8-nt site and the abundance of each site scaled by its normalized K_D. Abundance of 8-, 7-, and 6-nt, and low-affinity sites for miR-293 are plotted separately (purple, blue, cyan, and gray, respectively). The abundance of the canonical sites was determined by sequencing and that of the low-affinity sites was set such that the IC_{50} matched that observed in Figure 1C.

(B) Site occupancy for the simulations in (A). Plotted is the simulated free fraction of *mCherry* 1s reporter as a function of its expression, otherwise as in (A).

(C) Simulated effects of changing miR-293 concentrations in ESCs on 8-nt target site occupancy using the mathematical models of site competition from Bosson et al. (2014) (left) or Jens and Rajewsky (2015) (right). Simulated free fraction of an added 8-nt target site is plotted as a function of its expression as in (B), using the binding-affinity distributions of the simulated target sites of the original studies, plotted as in (A). The IC_{50} inferred from Figure 1C is indicated (dotted lines).

(D) Simulations using the models of (C) but adding low-affinity sites to alleviate the discrepancy between the simulated and experimental results, otherwise as in (C).

expected to be more easily observed and more biologically relevant when miRNA levels are high.

When we conclude that miRNA levels do not substantially influence the DRT, we refer to the DRT as the number of sites that were measured in steady-state conditions in the presence of miRNA-mediated repression, such as those represented by the *mCherry* transcripts in the dual-fluorescence reporter system. In this system, it was important to account for the miRNA-mediated degradation of the competitor, as our conclusions would have differed if we determined the competitor concentration in the absence of miRNA repression, represented by the output of the co-transcribed eYFP reporter.

Bulged and fully complementary sites were the first site types to be investigated in the context of regulating miRNA activity through competition (Ebert et al., 2007; Franco-Zorrilla et al., 2007), and they have been widely used to inhibit or measure miRNA activity (Doench et al., 2003; Broderick et al., 2011; Mukherji et al., 2011; Mullokandov et al., 2012; Xie et al., 2012). However, these sites with extensive complementarity to the 3′ region of the miRNA can trigger degradation of the miRNA (Ameres et al., 2010). Indeed, we observed target-directed miRNA degradation in hepatocytes when adding bulged sites of miR-122. Hence, bulged sites can reduce miRNA activity predominantly through triggering miRNA degradation rather than by competing with other miRNA-binding sites. Likewise, endogenous transcripts with highly complementary binding sites might affect miRNA activity through degradation rather than competition, especially in situations of low or intermediate

miRNA levels, with this degradation mechanism requiring much lower expression levels to be consequential. For example, potent target-directed degradation has been described in primary neurons (de la Mata et al., 2015), and a highly complementary binding site has been identified in the linc-MD1 long non-coding RNA and implicated in muscle differentiation through a ceRNA mechanism (Cesana et al., 2011). Whether this complementary site can induce miRNA degradation or whether other such sites exist remains to be shown.

We found that 7-nt sites were 50% as effective as 8-nt sites in contributing to target abundance, and 6-nt sites were 20% as effective. This 20% efficacy compared to 8-nt sites was much greater than might have been expected from the marginal repression typically imparted by 6-nt sites, again illustrating how competition efficacy imperfectly mirrors repression efficacy. Because miRNA association rates (k_{on} values) of 6-, 7-, or 8-nt sites are similar (Chandradoss et al., 2015; Salomon et al., 2015), the difference between competition and repression presumably relates to the different dissociation rates (k_{off} values) of these site types. Perhaps, before any repression can begin, some time is required to remodel the target transcript, assembling TNRC6 (trinucleotide repeat containing 6) and the deadenylation complexes, such that the dwell time of the miRNA on 6-nt sites only rarely exceeds this lag time. Similar models have been proposed to explain the poor repression efficacy of sites in the path of the ribosome (Grimson et al., 2007) and inefficacy of non-canonical sites in 3′ UTRs, despite the compelling CLIP evidence for binding to the ineffective sites (Agarwal et al., 2015). In this way, site types that are marginal or ineffective with respect to repression can nonetheless contribute meaningfully to effective target-site abundance. Indeed, our mathematical simulations illustrate that, in aggregate, low-affinity, non-canonical/background sites contribute more to the effective target-site abundance than do the canonical sites.

As the sum of 6-, 7-, and 8-nt sites in transcriptome 3′ UTRs, TA_{app} is a crude approximation of the effective target site abundance, in that it overcounts effects of 6- and 7-nt 3′ UTR sites and misses both the sites outside of 3′ UTRs and the weak but highly abundant non-canonical sites in 3′ UTRs. Nonetheless, summing up all 6-, 7-, or 8-nt 3′ UTR sites equally without weighting approximated IC_{50} within a few fold, presumably because overcounting the effects of some sites largely offset the failure to count other sites.

Our competition results provided mechanistic insight into the cooperative effect sometimes observed for adjacent sites. Whereas two distantly spaced 3′ UTR sites typically confer the repression expected from their independent action, two more closely spaced sites often confer more repression than expected from independent action (Grimson et al., 2007; Saetrom et al., 2007). Previous studies of this phenomenon using repression as the output do not distinguish between cooperative binding of the two sites or some other type of cooperative function in repression. Our use of competition as the output, with the observation that transcripts containing two miRNA-binding sites spaced 58 nt apart cooperatively sequester miRNAs corresponding to each site, uniquely shows that cooperative binding occurs.

Although the mechanism of this cooperative binding is unknown, an attractive hypothesis is that nearby Argonaute proteins might be tethered to each other through binding of the same TNRC6 molecule, also known as glycine-tryptophan protein of 182 kDa (GW182) in flies (Huntzinger and Izaurralde, 2011; Fabian and Sonenberg, 2012). TNRC6 contains multiple Argonaute-binding sites that might simultaneously interact with multiple miRNA-loaded Argonaute proteins (Schirle and MacRae, 2012; Pfaff et al., 2013), thereby enabling adjacent transcript-bound Argonaute proteins to prolong the dwell times of each other.

In the previous study of miR-122 site competition in hepatocytes, the three-site construct appears only 3-fold more effective than the one-site construct, as would be expected for non-cooperative, independent action of the three sites (Denzler et al., 2014). Suspecting that cooperativity was not observed in this context because the number of different MOIs examined was insufficient to detect subtle differences, we revisited potential cooperative binding of sites within the Ad-AldoA 3s construct at more MOIs. Derepression started to occur at 1.1×10^5 MREs (Figure S7), a DRT about 50% lower than that observed for the Ad-AldoA 1s construct. Thus, as expected, cooperativity can be observed for miRNAs of the same family as well as for miRNAs of different families.

Among the features that we analyzed, cooperative binding of miRNAs was the only one that increased the feasibility of regulation through changes in ceRNA levels, lowering the number of competing sites needed to detect derepression by ~50%. However, this 50% difference does not seem large enough to substantially improve the prospects of observing a ceRNA effect in a physiological setting. Perhaps in unusual cases cooperativity provides more than a 50% difference, a good candidate for unusually strong cooperativity being the circular RNA CDR1as/ciRS-7 with >60 closely spaced sites to miR-7 (Hansen et al., 2013; Memczak et al., 2013). However, very few other circRNAs have more miRNA-binding sites than expected by chance (Guo et al., 2014; Rybak-Wolf et al., 2015), which brings the focus back to linear transcripts as a more abundant source of potential ceRNA candidates. For cooperativity to be a factor, such a transcript would need to be very highly expressed and have multiple sites that fall in a cooperative sequence context, and sites would need to correspond to miRNA families that are each expressed at levels sufficient to actively repress target genes. If or how frequently such conditions occur in vivo is currently unknown, but if such a candidate is found, recently developed gene-editing methods offer the opportunity to introduce precise mutations of the sites within their genomic context (without induced overexpression) and thereby provide the first convincing evidence of ceRNA regulation in vivo.

EXPERIMENTAL PROCEDURES

See Supplemental Experimental Procedures for details.

Single-Cell Reporter Assay

The fluorescent reporter plasmids are based on the pTRE-Tight-BI (Clontech) system, in which a bidirectional Tet promoter expresses eYFP and mCherry (Mukherji et al., 2011). The 3′ UTR of *mCherry* contains either zero (0s), one (1s), or

three consecutive (3s) 48-nt-long sequence stretches, which are comprised of one 8-mer MRE and ±20-bp flanking regions (Bosson et al., 2014). ESC line E14 was transfected with reporter and rtTA plasmids, induced with doxycycline 6 hr post-transfection, and harvested 18 hr later. Samples were analyzed using a FACSAria IIIu flow cytometer and eYFP and mCherry fluorescent values were corrected for autofluorescence as described in Bosson et al. (2014).

Hepatocyte Isolation and Viral Infections

Animal experiments were approved by the Kantonale Veterinäramt Zürich. Hepatocytes of 8- to 12-week-old male C57BL/6N mice (Janvier) were counted and plated at 300,000 cells per well in surface-treated six-well plates (Corning) in low-glucose media. Four to 6 hr after plating, cells were infected with adenovirus constructs in Hepatozyme media (Life Technologies) and harvested 24 hr post-infection.

Adenoviruses

Recombinant adenoviruses generated in this study are based on the *AldoA* constructs described in Denzler et al. (2014) and express *GFP* from an independent promoter. See Tables S5 and S6 for the nucleotide sequences of all Ad-AldoA constructs.

miRNA and Gene Expression Analysis

qPCRs were performed using TaqMan MicroRNA Assays (Life Technologies) for miRNA or gene-specific primer pairs (Table S4) for gene expression, respectively. Relative expression values were calculated using the ddCT method employing snoRNA202 for miRNA or mouse *36b4* (*Rplp0*) for gene expression normalization.

ACCESSION NUMBERS

The accession number for the data reported in this paper is GEO: GSE76288.

SUPPLEMENTAL INFORMATION

Supplemental Information contains Supplemental Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2016.09.027.

AUTHOR CONTRIBUTIONS

R.D. designed and performed experiments, analyzed and interpreted data, and drafted the manuscript. S.E.M. developed the mathematical model and performed associated analyses. A.C.T. performed RNA blots and helped with experiments shown in Figure 1. V.A. processed RNA-seq raw data. M.S. and D.P.B. designed experiments, interpreted data, and revised the manuscript. R.D., S.E.M., A.C.T., V.A., D.P.B., and M.S. reviewed the results and contributed to writing the manuscript.

ACKNOWLEDGMENTS

We would like to thank J. Zamudio, C. JnBaptise, and P. Sharp for technical advice and helpful discussions; B. Kleaveland for small-RNA-seq; and C. Ciaudo for providing ESCs. This study was supported in part by the National Science Foundation Graduate Research Fellowship (V.A.), an ERC grant "Metabolomirs" and NCCR "RNA and Biology" (M.S.), and NIH grant GM067031 (D.P.B.). D.P.B. is a Howard Hughes Medical Institute Investigator.

Received: January 25, 2016
Revised: June 10, 2016
Accepted: September 20, 2016
Published: October 27, 2016

REFERENCES

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

Ala, U., Karreth, F.A., Bosia, C., Pagnani, A., Taulli, R., Léopold, V., Tay, Y., Provero, P., Zecchina, R., and Pandolfi, P.P. (2013). Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. Proc. Natl. Acad. Sci. USA *110*, 7154–7159.

Ameres, S.L., Horwich, M.D., Hung, J.H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P.D. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. Science *328*, 1534–1539.

Arvey, A., Larsson, E., Sander, C., Leslie, C.S., and Marks, D.S. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. Mol. Syst. Biol. *6*, 363.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bosson, A.D., Zamudio, J.R., and Sharp, P.A. (2014). Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. Mol. Cell *56*, 347–359.

Broderick, J.A., Salomon, W.E., Ryder, S.P., Aronin, N., and Zamore, P.D. (2011). Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. RNA *17*, 1858–1869.

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell *147*, 358–369.

Chandradoss, S.D., Schirle, N.T., Szczepaniak, M., MacRae, I.J., and Joo, C. (2015). A dynamic search process underlies microRNA targeting. Cell *162*, 96–107.

de la Mata, M., Gaidatzis, D., Vitanescu, M., Stadler, M.B., Wentzel, C., Scheiffele, P., Filipowicz, W., and Großhans, H. (2015). Potent degradation of neuronal miRNAs induced by highly complementary targets. EMBO Rep. *16*, 500–511.

Denzler, R., Agarwal, V., Stefano, J., Bartel, D.P., and Stoffel, M. (2014). Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. Mol. Cell *54*, 766–776.

Doench, J.G., Petersen, C.P., and Sharp, P.A. (2003). siRNAs can function as miRNAs. Genes Dev. *17*, 438–442.

Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods *4*, 721–726.

Fabian, M.R., and Sonenberg, N. (2012). The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. Nat. Struct. Mol. Biol. *19*, 586–593.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J.A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nat. Genet. *39*, 1033–1037.

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat. Struct. Mol. Biol. *18*, 1139–1146.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75–79.

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

Guo, J.U., Agarwal, V., Guo, H., and Bartel, D.P. (2014). Expanded identification and characterization of mammalian circular RNAs. Genome Biol. *15*, 409.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384–388.

Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat. Rev. Genet. *12*, 99–110.

Jens, M., and Rajewsky, N. (2015). Competition between target sites of regulators shapes post-transcriptional gene regulation. Nat. Rev. Genet. *16*, 113–126.

Karreth, F.A., Reschke, M., Ruocco, A., Ng, C., Chapuy, B., Léopold, V., Sjoberg, M., Keane, T.M., Verma, A., Ala, U., et al. (2015). The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell *161*, 319–332.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature *495*, 333–338.

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

Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A.D., Tung, N., Israelow, B., Evans, M.J., Sachidanandam, R., and Brown, B.D. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nat. Methods *9*, 840–846.

Pfaff, J., Hennig, J., Herzog, F., Aebersold, R., Sattler, M., Niessing, D., and Meister, G. (2013). Structural features of Argonaute-GW182 protein interactions. Proc. Natl. Acad. Sci. USA *110*, E3770–E3779.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol. Cell *58*, 870–885.

Saetrom, P., Heale, B.S., Snøve, O., Jr., Aagaard, L., Alluin, J., and Rossi, J.J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res. *35*, 2333–2342.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Salomon, W.E., Jolly, S.M., Moore, M.J., Zamore, P.D., and Serebrov, V. (2015). Single-molecule imaging reveals that Argonaute reshapes the binding properties of its nucleic acid guides. Cell *162*, 84–95.

Schirle, N.T., and MacRae, I.J. (2012). The crystal structure of human Argonaute2. Science *336*, 1037–1040.

Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Blüthgen, N., Marks, D.S., and van Oudenaarden, A. (2015). Gene expression. MicroRNA control of protein expression noise. Science *348*, 128–132.

Xie, J., Ameres, S.L., Friedline, R., Hung, J.H., Zhang, Y., Xie, Q., Zhong, L., Su, Q., He, R., Li, M., et al. (2012). Long-term, efficient inhibition of microRNA function in mice using rAAV vectors. Nat. Methods *9*, 403–409.

Yuan, Y., Liu, B., Xie, P., Zhang, M.Q., Li, Y., Xie, Z., and Wang, X. (2015). Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. Proc. Natl. Acad. Sci. USA *112*, 3158–3163.

Appendix E.

The dynamics of cytoplasmic mRNA metabolism

Timothy J. Eisen[1,2,3]*, Stephen W. Eichhorn[1,2,3]*, Alexander O. Subtelny[1,2,3]*, Kathy S. Lin[1,2,3,4], Sean E. McGeary[1,2,3], Sumeet Gupta[2], and David P. Bartel[1,2,3]

[1]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*These authors contributed equally to this work.

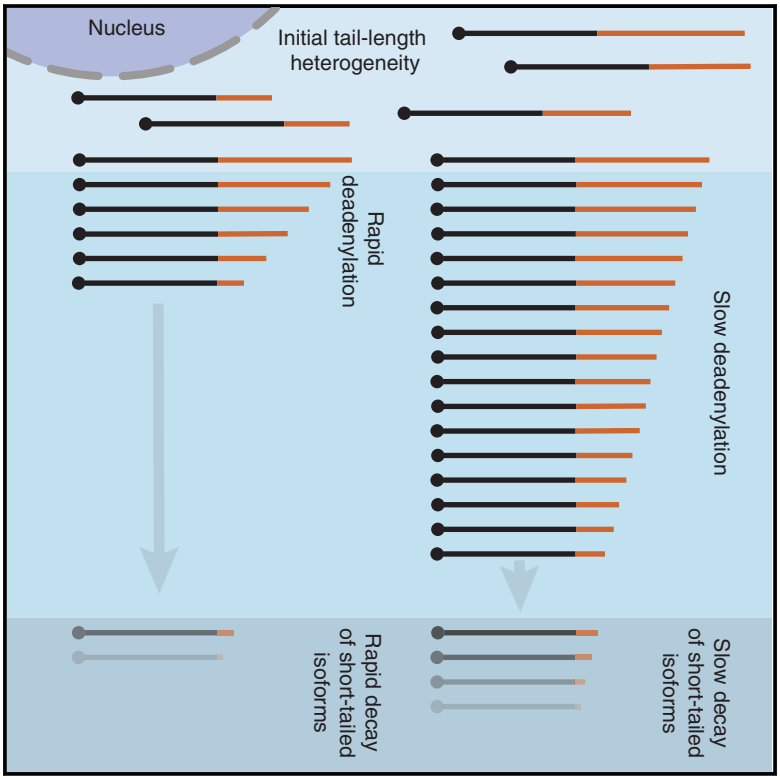
Published as:
Eisen, T.J., Eichhorn, S.W., Subtelny, A.O., Lin, K.S., McGeary, S.E., and Bartel., D.P. (2020). The dynamics of cytoplasmic mRNA metabolism. Mol. Cell *77*, 786–799.

Article

Molecular Cell

The Dynamics of Cytoplasmic mRNA Metabolism

Graphical Abstract



Highlights

- mRNAs enter the cytoplasm with diverse intra- and intergenic tail lengths
- mRNA deadenylation rates span a 1000-fold range and correspond to mRNA half-lives
- After their tails become short, mRNAs decay at rates that span a 1000-fold range
- More rapidly deadenylated mRNAs decay more rapidly upon reaching short tail lengths

Authors

Timothy J. Eisen, Stephen W. Eichhorn, Alexander O. Subtelny, Kathy S. Lin, Sean E. McGeary, Sumeet Gupta, David P. Bartel

Correspondence

dbartel@wi.mit.edu

In Brief

mRNA decay helps determine the extent of mRNA accumulation and ultimately the amount of protein produced. The dynamics of mRNA decay—involving tail-length shortening and then decay of the mRNA body—are largely unknown. Eisen et al. use high-throughput methods to uncover these dynamics for thousands of endogenous mRNAs.

Eisen et al., 2020, Molecular Cell 77, 786–799
February 20, 2020
https://doi.org/10.1016/j.molcel.2019.12.005

CellPress

CellPress

Molecular Cell
Article

The Dynamics of Cytoplasmic mRNA Metabolism

Timothy J. Eisen,[1,2,3,5] Stephen W. Eichhorn,[1,2,3,5] Alexander O. Subtelny,[1,2,3,5] Kathy S. Lin,[1,2,3,4] Sean E. McGeary,[1,2,3] Sumeet Gupta,[2] and David P. Bartel[1,2,3,6,*]

[1]Howard Hughes Medical Institute, Cambridge, MA 02142, USA
[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: dbartel@wi.mit.edu
https://doi.org/10.1016/j.molcel.2019.12.005

SUMMARY

For all but a few mRNAs, the dynamics of metabolism are unknown. Here, we developed an experimental and analytical framework for examining these dynamics for mRNAs from thousands of genes. mRNAs of mouse fibroblasts exit the nucleus with diverse intragenic and intergenic poly(A)-tail lengths. Once in the cytoplasm, they have a broad (1000-fold) range of deadenylation rate constants, which correspond to cytoplasmic lifetimes. Indeed, with few exceptions, degradation appears to occur primarily through deadenylation-linked mechanisms, with little contribution from either endonucleolytic cleavage or deadenylation-independent decapping. Most mRNA molecules degrade only after their tail lengths fall below 25 nt. Decay rate constants of short-tailed mRNAs vary broadly (1000-fold) and are larger for short-tailed mRNAs that have previously undergone more rapid deadenylation. This coupling helps clear rapidly deadenylated mRNAs, enabling the large range in deadenylation rate constants to impart a similarly large range in stabilities.

INTRODUCTION

mRNAs corresponding to different genes are degraded at substantially different rates, with some mRNAs turning over in minutes and others persisting for days (Dölken et al., 2008). Different conditions or developmental contexts can modify these rates, resulting in the destabilization of previously stable mRNAs, or vice versa (Rabani et al., 2011). These rate changes influence the dynamics of mRNA accumulation and, ultimately, the steady-state abundance of mRNAs.

Many proteins that promote mammalian mRNA degradation also can recruit deadenylase complexes. These include Pumilio (Van Etten et al., 2012), SMG5/7 (Mühlemann and Lykke-Andersen, 2010), GW182 (Fabian et al., 2011), BTG/TOB factors (Mauxion et al., 2009), Roquin (Leppek et al., 2013), YTHDF2 (Du et al., 2016), and HuR, TTP, and other proteins that bind AU- and GU-rich elements (Vlasova-St Louis and Bohjanen, 2011; Fabian et al., 2013). That these diverse modifiers of mRNA stability converge on deadenylation suggests that differences in deadenylation rates might explain a substantial fraction of the variation observed in mRNA stability.

In the past, the dynamics of mRNA deadenylation have been examined on a gene-by-gene basis, involving pulsed expression and subsequent analysis of mRNA transcripts using RNase H to cleave the mRNA and RNA blots to probe for the poly(A)-tailed 3′ fragment. Because this procedure has been performed for only a handful of cellular mRNAs in yeast (Decker and Parker, 1993; Muhlrad et al., 1994; Hilgers et al., 2006) and mammals (Mercer and Wake, 1985; Wilson and Treisman, 1988; Shyu et al., 1991; Chen and Shyu, 1995; Gowrishankar et al., 2005), some fundamental questions, including the extent to which a global relationship exists between deadenylation rate and mRNA stability, have remained unanswered.

Here, we developed experimental and analytical tools for the global analysis of tail-length dynamics. Applying these tools to the mRNAs of cultured mouse fibroblasts generated a unique resource of initial cytoplasmic tail lengths, deadenylation rates, and decay parameters for mRNAs of thousands of individual genes, which in turn provided fundamental insights into cytoplasmic mRNA metabolism.

RESULTS

Global Profiling of Tail-Length Dynamics

Two high-throughput methods, each with distinct advantages, were initially developed to profile poly(A)-tail lengths. One is PAL-seq (poly(A)-tail-length profiling by sequencing), which also reports the cleavage-and-polyadenylation site for each polyadenylated molecule (Subtelny et al., 2014), whereas the other is TAIL-seq, which can measure poly(A) tails that have been terminally modified with non-A residues (Chang et al., 2014; Lim et al., 2016). Here, we developed PAL-seq version 2 (v2), which combines these advantages and has the further benefit over both previous methods of more robust compatibility with contemporary Illumina sequencing platforms (Figure S1).

To observe tail-length dynamics of endogenous mRNAs, we employed a metabolic-labeling approach in which mRNAs of different age ranges were isolated and analyzed (Figure 1A). To initiate labeling, we added 5-ethynyl uridine (5EU) to 3T3 cells.

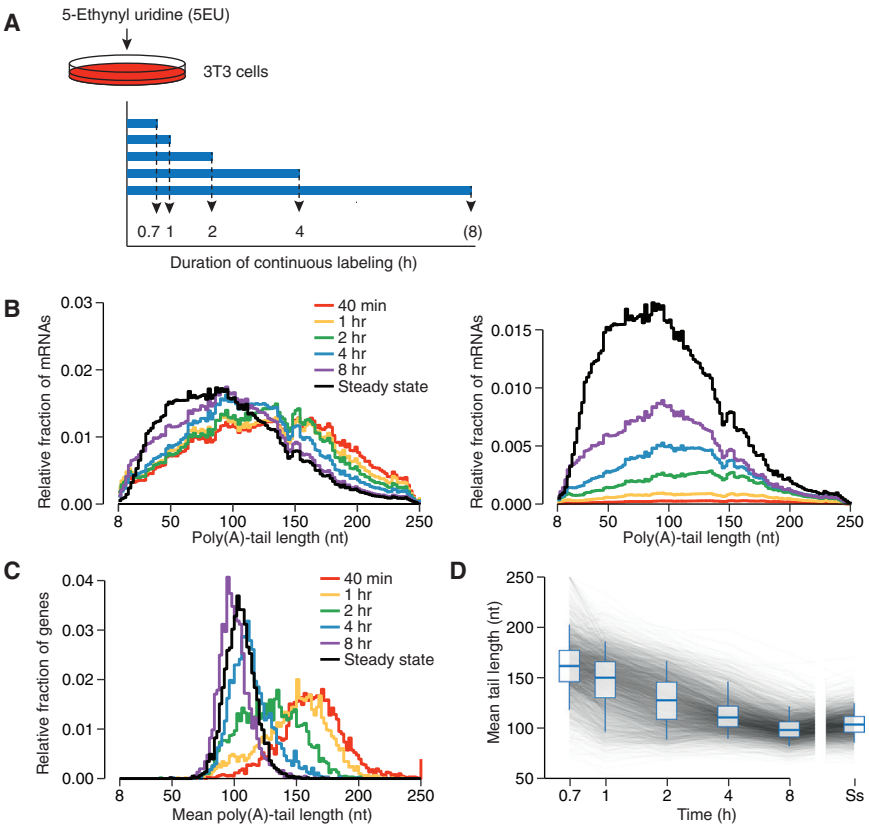


Figure 1. Global Tail-Length Dynamics of Mammalian mRNAs

(A) Schematic of 5EU metabolic labeling. Experiments were performed with two 3T3 cell lines designed to induce expression of either miR-155 or miR-1 (cell lines 1 and 2, respectively) but cultured without microRNA induction. The 8 is in parentheses because an 8-h labeling period was included for only one line (cell line 1). For simplicity, all subsequent figures show the results for cell line 1, unless stated otherwise.

(B) Tail-length distributions of mRNA molecules isolated after each period of 5EU labeling (key). Left: distributions were normalized to each have the same area. Right: distributions were scaled to the abundance of labeled RNAs in each sample and then normalized such that the steady-state sample had an area of 1. The steady-state sample was prepared with unselected RNA from the 40-min time interval. Each bin is 2 nt; results for the bin with tail lengths $\geq$ 250 nt are not shown.

(C) Distributions of mean poly(A)-tail lengths for mRNAs of each gene after the indicated duration of 5EU labeling. Values for all genes that passed the tag cutoffs for tail-length measurement at all time intervals were included (n = 3,048). Each bin is 2 nt. Genes with mean mRNA tail-length values greater than $\geq$ 250 nt were assigned to the 250-nt bin.

(D) Tail lengths over time. Mean tail lengths for mRNAs from each gene (n = 3,048) are plotted along with box-and-whiskers overlays (line, median; box, 25th–75th percentiles; whiskers, 5th–95th percentiles). Ss, steady state. See also Figures S1 and S2.

After incubating for time periods ranging from 40 min to 8 h, cytoplasmically enriched lysates were collected, and RNA containing 5EU was isolated by virtue of the reactivity between the 5EU and an azide-bearing biotin tag. Poly(A)-tail lengths of captured mRNAs, as well as total-lysate mRNA, were measured using PAL-seq v2 (hereafter called PAL-seq). In parallel, we performed RNA-seq, which measured mRNA abundance for each time interval. Spike-in of RNA standards with known tail lengths enabled estimates of recovery and measurement accuracy over a broad range of tail lengths, as well as absolute quantification of RNA measured by each method. These experiments were performed using each of two independently passaged 3T3 cell lines. Unless stated otherwise, figures show the results obtained for cell line 1. Nonetheless, the results of the two cell lines were reproducible at each time interval ($R_s \geq 0.81$ for mean tail-length measurements). Moreover, results from either PAL-seq v1, PAL-seq v2, or our implementation of TAIL-seq were highly correlated (Figures S2A–S2D; R_s = 0.83–0.88 for each of the two-way comparisons), which indicated that our conclusions were independent of the method used for tail-length profiling.

As expected if tail lengths become shorter over time in the cytoplasm (Sheiness and Darnell, 1973; Palatnik et al., 1979), mRNAs collected after the shortest labeling period (40 min) had the longest poly(A)-tail lengths, with median length of 133 nt (Figure 1B). As the average age of each labeled mRNA population increased with longer labeling periods, tail-length distributions shifted toward the steady-state distribution with respect to both length and abundance (Figure 1B). At each time interval, 10–20-nt tails preferentially possessed a 3′ terminal U (Figure S2E), although < 6.8% of tails had 3′ U residues in any sample, in keeping with previous reports on the fraction of short tails with terminal uridines at steady state (Chang et al., 2014; Lim et al., 2014). Analyses of mean poly(A)-tail lengths for mRNAs corresponding to thousands of individual genes showed that tails from mRNAs of essentially every gene shortened over time in the cytoplasm (Figures 1C and 1D).

Correspondence between mRNA Half-Life and Deadenylation Rate

After 2 h of labeling, a broad range of mean tail lengths was observed, as mean tail lengths for mRNAs of some genes approached their steady-state values, whereas those for others still resembled their initial values (Figure 1C). These different rates of approach to steady state presumably at least partly reflected differences in mRNA degradation rates, as short-lived mRNAs were expected to reach their steady-state abundance and poly(A)-tail length more rapidly than long-lived mRNAs.

To determine these degradation rates, we fit the yield of PAL-seq tags obtained for each gene at each time interval (normalizing to the spike-in controls) to the exponential function describing the approach to steady state, while also fitting a global offset to account for a delay between the time that 5EU was added and the time that labeled mRNAs appeared in the cytoplasm. This offset ranged from 27 to 36 min, depending on the experiment, a range

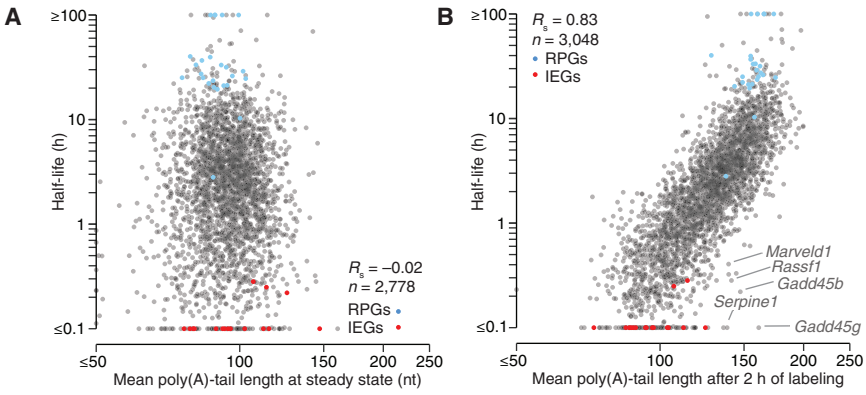


Figure 2. Correspondence between mRNA Half-Life and Deadenylation Rate

(A) Relationship between half-life and mean steady-state tail length of mRNAs in 3T3 cells. For mRNAs of each gene, standard PAL-seq data were used to determine the length distribution of tails ≥ 50 nt, and data generated from a protocol that used single-stranded ligation to the mRNA 3′ termini (rather than a splinted ligation to the tail) were used to determine both the length distribution of tails < 50 nt and the fraction of tails < 50 nt. Compared to the tail-length distribution generated by only standard PAL-seq data, this composite distribution better accounted for very short and highly modified tails. Nonetheless, using the standard PAL-seq data without this adjustment produced a similar result (Figure S3G). Results for mRNAs of ribosomal protein genes (RPGs) and immediate-early genes (IEGs) (Tullai et al., 2007) are indicated (blue and red, respectively).

(B) Relationship between mRNA half-life and mean tail length of metabolically labeled mRNAs isolated after 2 h of labeling. Otherwise as in (A).

See also Figures S3A–S3D and S3G.

consistent with single-gene measurements of the time required for mRNA transcription, processing, and export (Shav-Tal et al., 2004; Mor et al., 2010). Our half-life values (Table S1) correlated well with those previously reported for mRNAs of 3T3 cells growing in similar conditions (Schwanhäusser et al., 2011) (Figure S3A; R_s = 0.68–0.77), although our absolute values were substantially shorter (Figures S3B–S3D; median 2.1 h for mRNAs of the 3T3 cell line 1, as opposed to 9 h for previously reported values). This difference was attributable to potential divergence in the cell lines used in the two labs, as well as our focus on cytoplasmically enriched RNA and our absolute quantification of labeled RNA (enabled by spiking in standards).

Previous global analyses of the relationship between mRNA half-life and mean tail length have been limited to steady-state tail-length measurements, for which no positive relationship is observed (Subtelny et al., 2014), despite the established role of poly(A) tails in conferring mRNA stability. Our current datasets, which provided the opportunity to make this comparison using half-life and tail-length measurements acquired from the same cells, reinforced this finding; we observed no positive relationship between mRNA half-life and mean steady-state tail length (Figure S3G; R_s = −0.24). This result held when incorporating results of PAL-seq implemented with direct ligation to mRNA 3′ termini, which better detected very short or highly modified tails (Figure 2A; R_s = −0.02). Indeed, the mean tail lengths of long-lived mRNAs, including those of ribosomal protein genes (RPGs), closely resembled tail lengths of short-lived mRNAs, including those of immediate-early genes (IEGs) (Figure 2A).

A very different picture emerged when considering pre-steady-state tail-length measurements. After 2 h of labeling, half-life strongly corresponded to mean tail length (Figure 2B; R_s = 0.83). At this labeling interval, IEG mRNAs and other short-lived mRNAs had the shortest mean tail lengths, RPG mRNAs and other long-lived mRNAs had the longest mean tail lengths, and other mRNAs had mean tail lengths falling somewhere in between. The simplest explanation for this result is that the deadenylation rate dictates the stability of most mRNAs, and mean tail length at 2 h provides a proxy for deadenylation rate. Thus, slow deadenylation of long-lived mRNAs explains both why they have longer tails after 2 h of labeling and why they have such long half-lives, and rapid deadenylation of short-lived mRNAs explains why they have shorter tails after 2 h of labeling and why they have such short half-lives.

Several notable outliers had half-lives that were shorter than expected from their mean tail lengths in the 2-h sample, suggesting that their degradation and deadenylation rates were incongruous. *Rassf1*, *Serpine1*, and two *Gadd45* paralogs are known or suspected substrates for either nonsense-mediated decay (NMD) or other pathways that recruit UPF1 (Nelson et al., 2016; Park and Maquat, 2013; Tani et al., 2012). Another outlier, the *Marveld1* mRNA, has not yet been reported to interact with UPF1, but its protein product does interact with UPF1 in human cells and regulates UPF1 activity (Hu et al., 2013). Association with UPF1 can trigger endonucleolytic cleavage of mammalian mRNAs, which would decouple the rates of decay and deadenylation (Mühlemann and Lykke-Andersen, 2010), disrupting the relationship between half-life and tail length at intermediate labeling intervals. Nonetheless, the most notable feature of the outliers was their scarcity; the striking overall correspondence observed between half-life and mean tail lengths after 2 h of labeling implied that for the vast majority of endogenous mRNA molecules of mouse fibroblasts, the rate of mRNA deadenylation largely determines the rate of degradation.

Initial Tail Lengths of Cytoplasmic mRNAs

Analysis of tail-length distributions for individual genes and the changes in these distributions over increased labeling intervals supported and extended the conclusions drawn from global analyses of abundances and mean tail lengths. This analysis confirmed that tail-length dynamics of mRNAs with short half-lives (e.g., *Metrnl*) substantially differed from those of mRNAs with longer half-lives (e.g., *Lsm1* and *Eef2*), with the short-lived mRNAs reaching their steady-state abundance and tail-length distribution much more rapidly (Figure 3). The stacked pattern of the distributions observed over increasing time intervals also illustrated that the longest-tailed mRNAs observed at steady state were essentially all recently transcribed, whereas the shortest-tailed mRNAs were mostly the oldest mRNAs (Figure 3).

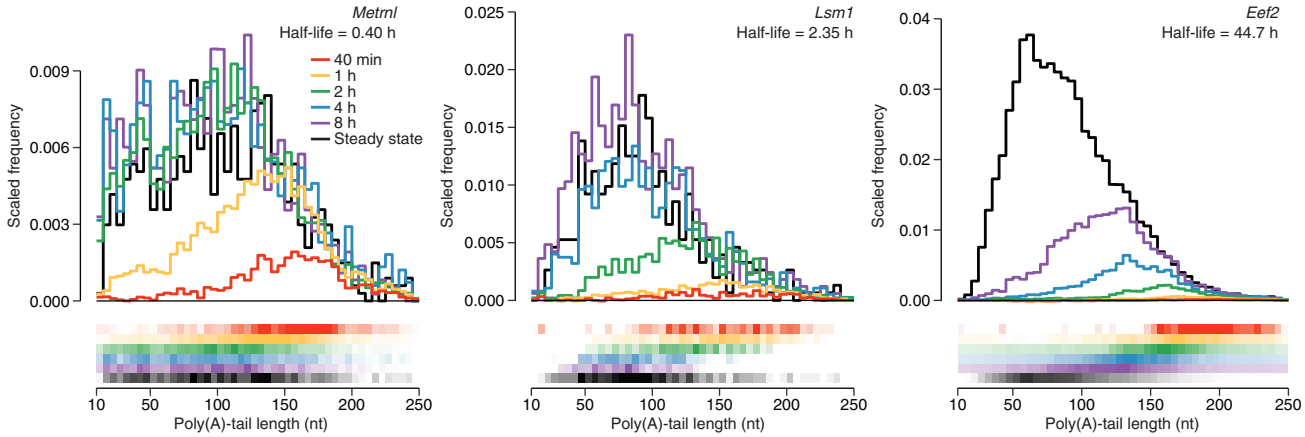


Figure 3. Tail-Length Dynamics of mRNAs with Different Half-Lives

Tail-length distributions for mRNAs from individual genes. For each time interval (key), the distribution is scaled to the abundance of labeled RNA in the sample (top), and the distribution is also represented as a heatmap (bottom), with the range of coloration corresponding to the 5th–95th percentiles of the histogram density. Each bin is 5 nt. Bins for tails < 10 nt are not shown because the splinted ligation to the tail used in the standard PAL-seq protocol depletes measurements for tails < 8 nt. Bins for tails ≥ 250 nt are also not shown.

See also Figures S3F and S3H–S3J.

Our tail-length data from short labeling periods provided the opportunity to examine the initial tail lengths of mRNAs soon after they entered the cytoplasm. The calculated 27–36-min delay in the appearance of labeled cytoplasmic mRNAs implied that most mRNAs isolated after 40 min of labeling were subject to cytoplasmic deadenylation for < 13 min. Thus, for all but the most rapidly deadenylated mRNAs, the tail lengths observed after 40 min of labeling should have approximated the tail lengths of mRNAs that first entered the cytoplasm.

Without data to the contrary, previous studies of tail-length dynamics have assumed that initial cytoplasmic tail lengths observed for mRNAs of one gene also apply to the mRNAs of all other genes. However, we observed substantial intergenic variation for average tail lengths at the shortest labeling period (Figures 1C, 3, and S3F), with the spread of the 5th–95th percentile values at least that of steady state (112.2 ± 4.7 to 194.7 ± 6.0 nt for the 40-min samples and 84.8 ± 1.3 to 124.6 ± 2.1 nt for the steady-state samples; values ± SD), which suggested that mRNAs from different genes exit the nucleus with tails of quite different lengths. To examine whether deadenylation occurring soon after nucleocytoplasmic export might have influenced this result, we focused on mRNAs with half-lives > 8 h. On average, mean tail lengths for these genes exhibited less than 4% change when comparing the 40-min and 1-h time intervals, implying that they also underwent little cytoplasmic deadenylation during the first 40 min of labeling. Average tail lengths observed at 40 min for mRNAs from these genes spanned a broad range, exceeding that observed at steady state (spread of the 5th–95th percentile values 128.3 ± 5.2 to 242.1 ± 16.1 nt for the 40-min samples and 81.0 ± 1.0 to 119.4 ± 1.4 nt for the steady-state samples; values ± SD), although these tail-length values observed at 40 min had little correspondence with those observed at steady state (R_s = 0.12).

When comparing mRNAs from the same gene, tail-length distributions were also quite broad for the newly exported mRNAs, as illustrated for mRNAs from three genes (Figure 3), and further demonstrated by the mean coefficient of variation (c.v.) of 0.41 for mRNAs of all measured genes (Figure S3H), compared to a c.v. of 0.20 for the 160-nt standard spiked into the 40-min sample. These c.v. values were reproducible between biological replicates and had little correspondence with mRNA half-life (Figures S3I and S3J). Although we cannot rule out the formal possibility that mRNA tails undergo exceedingly rapid and variable transient deadenylation immediately upon nuclear export, we interpret our results at short labeling periods to indicate that mRNAs exit the nucleus with considerable but reproducible intergenic and intragenic tail-length variability.

A Quantitative Model of mRNA Deadenylation and Decay

Our ability to isolate mRNAs of different age ranges for each gene and analyze their abundances and tail lengths (Figure 3) provided the unique opportunity to calculate the deadenylation rates and other metabolic rates and parameters for these mRNAs, thereby expanding the number of metabolically characterized mammalian mRNAs far beyond the four (*Mt1*, *Fos*, *Hbb*, and *IL8*) that have been examined using single-gene measurements (Mercer and Wake, 1985; Wilson and Treisman, 1988; Shyu et al., 1991; Gowrishankar et al., 2005). For each gene, the number of mRNA molecules with a given tail length is a function of (1) the rate of mRNA entering the cytoplasm, which in turn is a function of the rates of transcription, processing, and nucleocytoplasmic export; (2) the tail-length distribution of mRNA entering the cytoplasm; (3) the deadenylation rate; (4) the tail length below which the mRNA body is no longer protected from decay; and (5) the decay rate of the mRNA body (presumably preceded by decapping). Therefore, we developed a mathematical model to determine, for mRNAs from thousands of genes, values for each of these parameters.

Our model was based on a system of differential equations that describe the rates of change of abundance of mRNA intermediates (Figure 4A; Table S2), an approach resembling that

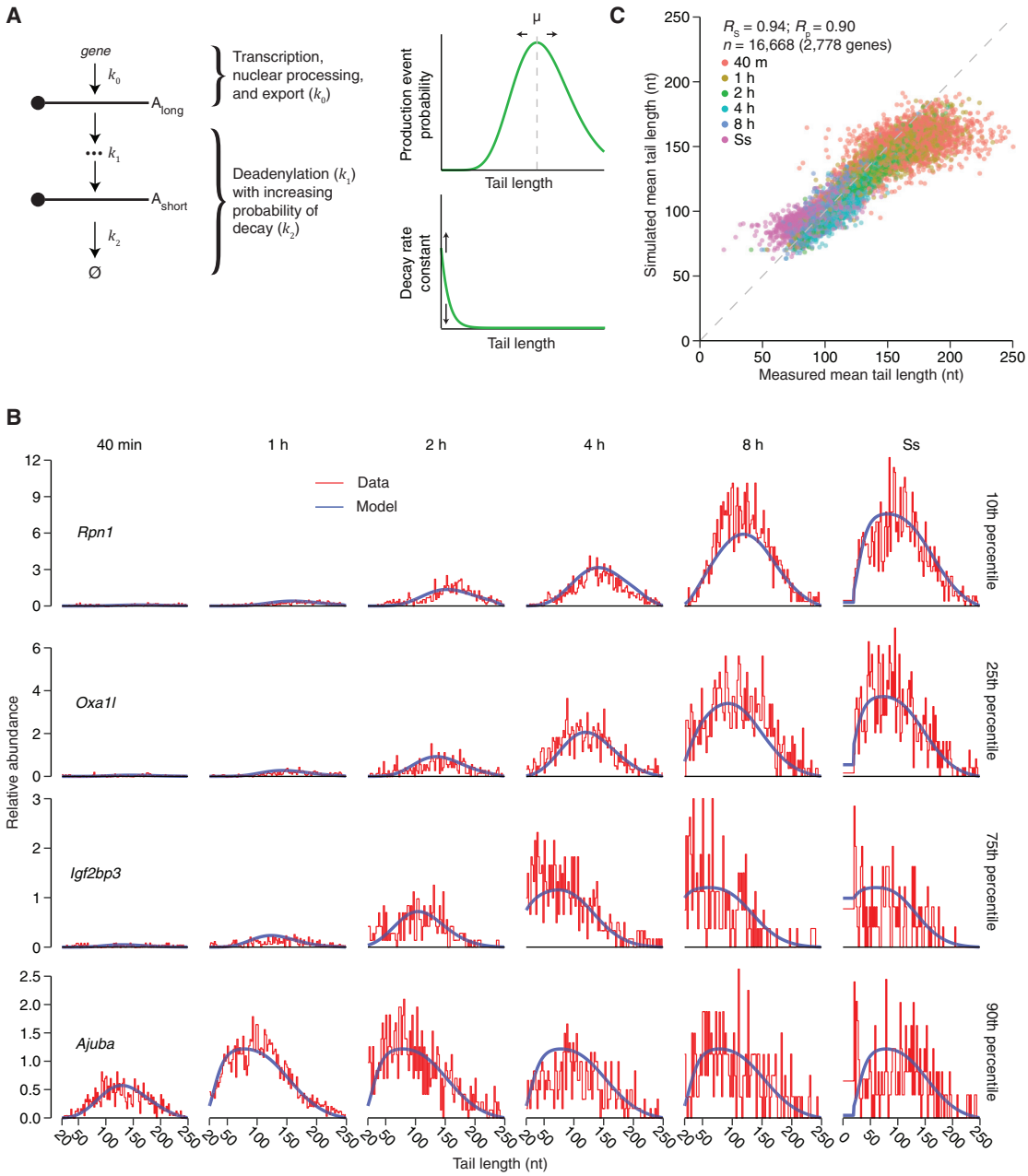


Figure 4. Computational Model of mRNA Deadenylation and Decay Dynamics

(A) Schematic of the computational model. k_0, k_1, and k_2 are terms for mRNA production, deadenylation, and decay, respectively, and ∅ represents the loss of the mRNA molecule. The curves (right) indicate the distributions used to model probabilities of production and decay as functions of tail length. They are schematized using the globally fitted parameters (v_p, m_d, and v_d) that defined each distribution (Table S2). The parameter m_p controls the mean (μ) of the negative binomial distribution (top curve), whereas the decay rate constant, β, scales the decay distribution (bottom curve) (Table S2).

(B) Correspondence between the model and the experimental data. Results for mRNAs of these four genes are shown as representative examples because their fits fell closest to the 10th, 25th, 75th, and 90th percentiles of the distribution of R^2 values for all genes that passed expression cutoffs in the PAL-seq datasets (Figure S4F; n = 2,778). For each time interval, the blue line shows the fit to the model, and the red line shows the distribution of observed tail-length species, plotted in 2-nt bins and scaled to standards as in Figure 1B, right. Ss, steady state.

(C) Correspondence between mean tail lengths generated from the model simulation and tail lengths measured in the metabolic labeling experiment. Shown for each gene are mean tail lengths for mRNAs at each time interval (key) from the simulation plotted with respect to the values observed experimentally. The discrepancy observed for some mRNAs at early time intervals was attributable to low signal for long-lived mRNAs at early times. The dashed line indicates $y = x$. See also Figures S4, S5A, and S5B and Tables S1 and S2.

used to model the metabolism of RNAs from single-gene reporters (Cao and Parker, 2001; Jia et al., 2011). For each gene, transcription, nuclear processing, and export (hereafter abbreviated as "production") generates, with rate constant k_0, a distribution of initial poly(A)-tail lengths. Over time, deadenylation shortens the tail, one nucleotide at a time, with rate constant k_1. Decay of the mRNA body, with rate constant k_2, can occur alongside deadenylation and monotonically increases as the poly(A) tails get shorter. One interpretation of this deadenylation-dependent decay is that it represents decapping, followed by rapid degradation of the mRNA body. However, because we do not monitor cap status, our model was not designed to distinguish between decay mechanisms and is compatible with either 5′ or 3′ exonucleolytic decay of the mRNA body.

For individual mRNAs generated from the same gene, the production terms varied according to a negative binomial distribution—a distribution routinely used to model the probability of a failure after a series of successes (in our case, creating an mRNA of tail length $n + 1$ after successfully creating an mRNA of tail length n) (Figure 4A; Table S2). The decay rate constant (k_2) followed a logistic function, which accelerated as tails shortened. The two parameters of this function (m_d and v_d) were fit as global constants, while the scaling parameter (β) was fit to each gene (Table S2). Solving the differential equations of the model estimated both the tail-length distribution and the mRNA abundance at each time interval for mRNAs from each gene.

Before arriving at the final version of the model (Figure 4A), we considered alternative models with varying levels of complexity. For example, building on the proposal that most mRNAs are substrates for both the PAN2/PAN3 and CCR4/NOT deadenylase complexes, with PAN2/PAN3 acting on tails > 110 nt and CCR4/NOT acting on shorter tails (Yamashita et al., 2005), we tested the performance of a model with two deadenylation rate constants, in which the transition between the two occurred at a tail length of 110 nt (Figure S4A). This model yielded residuals that were only marginally improved (Figure S4B), and for each mRNA the two deadenylation rates resembled each other (Figure S4C). A model in which the transition between the deadenylation rates occurred at 150 nt (Yi et al., 2018) yielded similar results (Figures S4D and S4E). These results indicated that, for endogenous mRNAs in 3T3 cells, either a single deadenylase complex dominates—as recently proposed for mRNAs with tail lengths $\leq$ 150 (Yi et al., 2018)—or both complexes act with indistinguishable kinetics. Thus, we chose not to implement a more complex model with two deadenylation rate constants.

Fitting the final version of the model to the tail-length and abundance measurements for mRNAs from thousands of genes yielded average initial tail lengths and rate constants for production, deadenylation, and deadenylation-dependent decay for each of these mRNAs (Table S2). The correspondence between the output of the model and the experimental measurements is illustrated for genes selected to represent different quantiles of fit based on the distribution of R^2 values (Figure 4B; Figure S4F). Mean tail-length values generated by the model corresponded well to measured values (Figure 4C; $R_s = 0.94$, $R_p = 0.90$). Moreover, values fit for starting tail length, production, deadenylation, and deadenylation-dependent decay were reproducible between biological replicates and robust to parameter initialization as well as multinomial sampling (bootstrap analysis) (Figures S4G–S4J).

The Dynamics of Cytoplasmic mRNA Metabolism

Of the six yeast mRNAs and four mammalian mRNAs that have been metabolically characterized, the data for four yeast mRNAs and two mammalian mRNAs are of sufficient resolution to derive deadenylation rates. The two mammalian mRNAs, *Fos* and *Mt1*, have deadenylation rate constants that differ by 60-fold (20 and 0.33 nt/min, respectively) (Mercer and Wake, 1985; Shyu et al., 1991). Our analysis, which characterized the metabolism of 2,778 mRNAs, greatly expanded the set of mRNAs with measured deadenylation rates and showed that deadenylation rate constants of mammalian mRNAs can differ by > 1000-fold—as fast as > 30 nt/min and as slow as 1.8 nt/h (Figure 5A). Concordant with our direct analysis of the primary data, which indicated that most mRNAs degrade through a mechanism involving tail shortening (Figure 1F), mRNA half-lives corresponded strongly to deadenylation rate constants fit to our model ($R_s = -0.95$; Figure S5A).

Our model and its fitted parameters allowed us to compute the deadenylation-dependent decay rates at each tail length and thereby infer the tail lengths at which mRNAs were degraded (Figure 5B). This analysis indicated that nearly all decay of the mRNA body occurred after the tail lengths fell below 100 nt, which agreed with previous analyses of reporter genes (Yamashita et al., 2005). Decay accelerated as tail lengths fell below 50 nt (with > 92% of mRNAs decaying below this length), a length less than the 54-nt footprint of two adjacent cytoplasmic poly(A)-binding protein (PABPC) molecules (Baer and Kornberg, 1983; Yi et al., 2018), but most mRNA molecules (> 55%) did not decay until their tail lengths fell below 25 nt, a length less than the 27-nt footprint of a single PABPC molecule (Figure 5B).

When analyzing for mRNAs of each gene the mean tail length at which the mRNA body decays, the results generally concurred with those observed for all mRNAs combined, with mRNAs from most genes decaying at short mean tail lengths (Figure 5C; > 97% decaying at mean tail length < 50 nt and > 69% decaying at mean tail length < 25 nt). As expected, most mRNAs previously found to have discordant deadenylation and decay rates (Figure 1F) were also outliers in this analysis, with *Gadd45b* and *Marveld1* degrading at mean tail lengths of 62, and 59 nt, respectively. The estimates of mean tail lengths at which mRNAs decay together with initial tail lengths and deadenylation rate constants enabled estimates of the time required to reach the mean tail length of decay, which corresponded to lifetime slightly better than did the deadenylation rate constants on their own to half-life (Figures S5A and S5B; $R_s = -0.96$ and -0.95, respectively.)

Once tails reached a short length, the decay rate constants varied widely, with short-tailed mRNAs from some genes undergoing decay at rate constants > 1000-fold greater than those of short-tailed mRNAs from other genes (Figure 5D). *Fos*, a rapidly deadenylated mRNA, is degraded much faster upon reaching a short tail length than is *Hbb*, a less rapidly deadenylated mRNA (Shyu et al., 1991). More rapid degradation of short-tailed mRNAs that had been more rapidly deadenylated would help prevent the buildup of short-tailed isoforms of rapidly

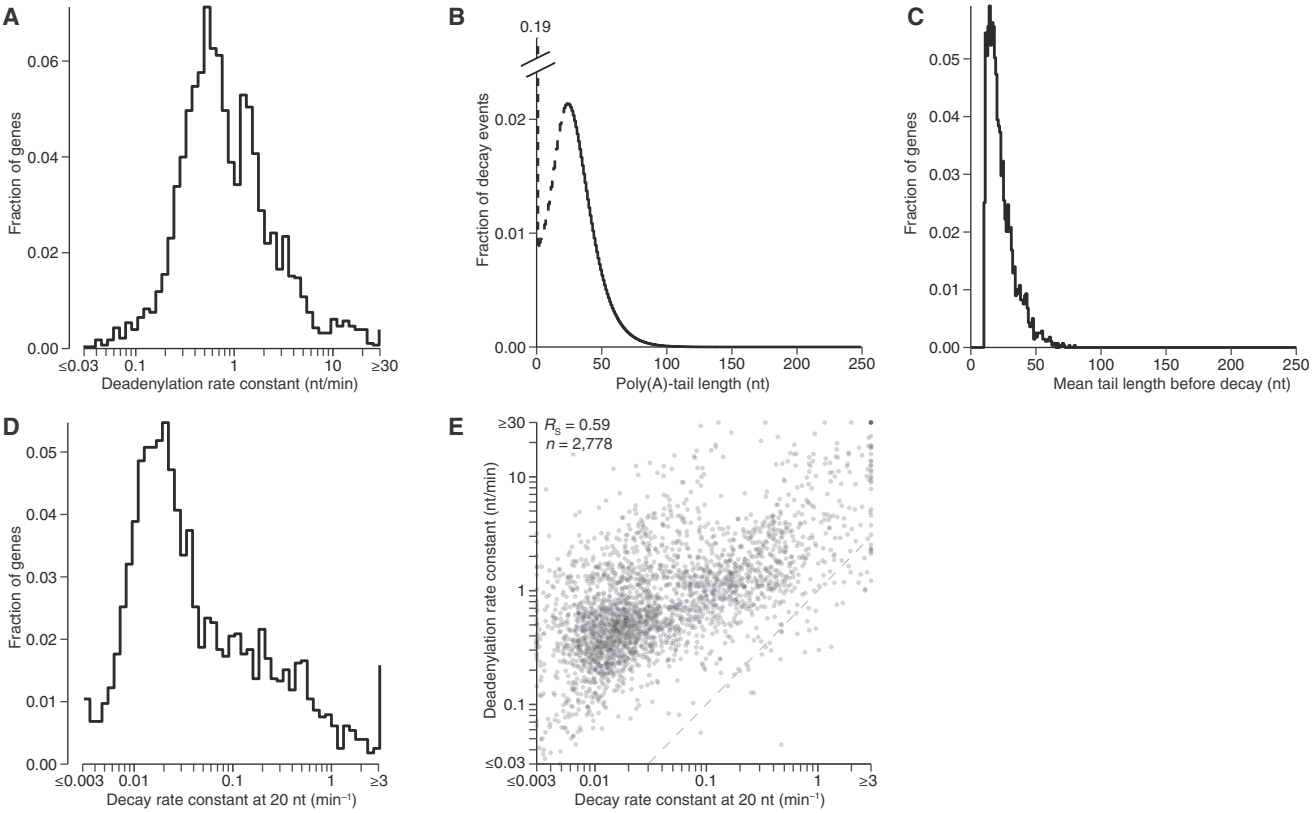


Figure 5. Dynamics of Cytoplasmic mRNA Metabolism

(A) Distribution of deadenylation rate constants (k_1 values), as determined by fitting the model to data for mRNAs from each gene (n = 2,778).

(B) Tail lengths at which mRNAs decay, as inferred by the model. The model rate constants were used to simulate a steady-state tail-length distribution for each gene. The abundance of each mRNA intermediate was then multiplied by the decay rate constant k_2 to yield a distribution of decay events over all tail lengths. Plotted is the combined distribution for all mRNA molecules of all 2,778 genes. Results were indistinguishable when the distribution from each gene was weighted equally. Values for tails < 20 nt are shown as a dashed line because the model fit steady-state tail lengths < 20 nt as an average of the total abundance of tails in this region and, thus, did not provide single-nucleotide resolution for decay rates of these species.

(C) Mean tail lengths at which mRNAs from each gene (n = 2,778) decayed, as inferred by the model. Otherwise, as in (B).

(D) Distribution of decay rate constants (k_2 values) for mRNAs with 20-nt tail lengths, as determined by fitting the model to data for mRNAs from each gene (n = 2,778).

(E) Correlation between the deadenylation rate constant (k_1) and the decay rate constant (k_2) at a tail length of 20 nt. The dashed line indicates $y = x$.

See also Figure S4.

deadenylated mRNAs. However, such buildup sometimes does occur, as observed in *Drosophila* cells for three mRNAs characterized during heat shock (Dellavalle et al., 1994; Bönisch et al., 2007) and in mammalian cells for *Csf2* (Chen et al., 1995; Carballo et al., 2000), raising the question of the extent to which decay rates of short-tailed mRNAs are coupled to their deadenylation rates. To answer this question, we examined the relationship between rate constants for deadenylation and those for decay of short-tailed mRNAs (the latter calculated for mRNAs with 20-nt tails). We found that more rapidly deadenylated mRNAs tended to be degraded more rapidly upon reaching short tail lengths (Figure 5E; R_s = 0.59).

A Modest Buildup of Short-Tailed Isoforms of Short-Lived mRNAs

Having found a strong tendency for more rapid clearing of mRNAs that had been more rapidly deadenylated, we investigated whether this phenomenon was able to prevent a large buildup of short-tailed isoforms of rapidly deadenylated mRNAs. For this investigation, we analyzed the steady-state dataset that incorporated results of PAL-seq implemented with direct ligation to mRNA 3′ termini, which better detected very short or highly modified tails. Despite the rapid decay of short-tailed mRNAs that had been more rapidly deadenylated, less-stable mRNAs generally did have a somewhat higher fraction of short-tailed transcripts (Figures 6A and S5C; R_s = −0.56). Nonetheless, the buildup of short-tailed isoforms of these unstable RNAs usually failed to exceed 30% of all transcripts (Figure 6A).

This preferential buildup of short-tailed isoforms of unstable RNAs was more clearly visualized in a meta-transcript analysis of the tail-length distribution at steady state. Short-lived mRNAs (half-lives < 20 min) had two peaks of short-tailed isoforms, a major peak centering at 7–15 nt and a minor peak at 0–1 nt, whereas long-lived mRNAs (half-lives > 10 h) were depleted of

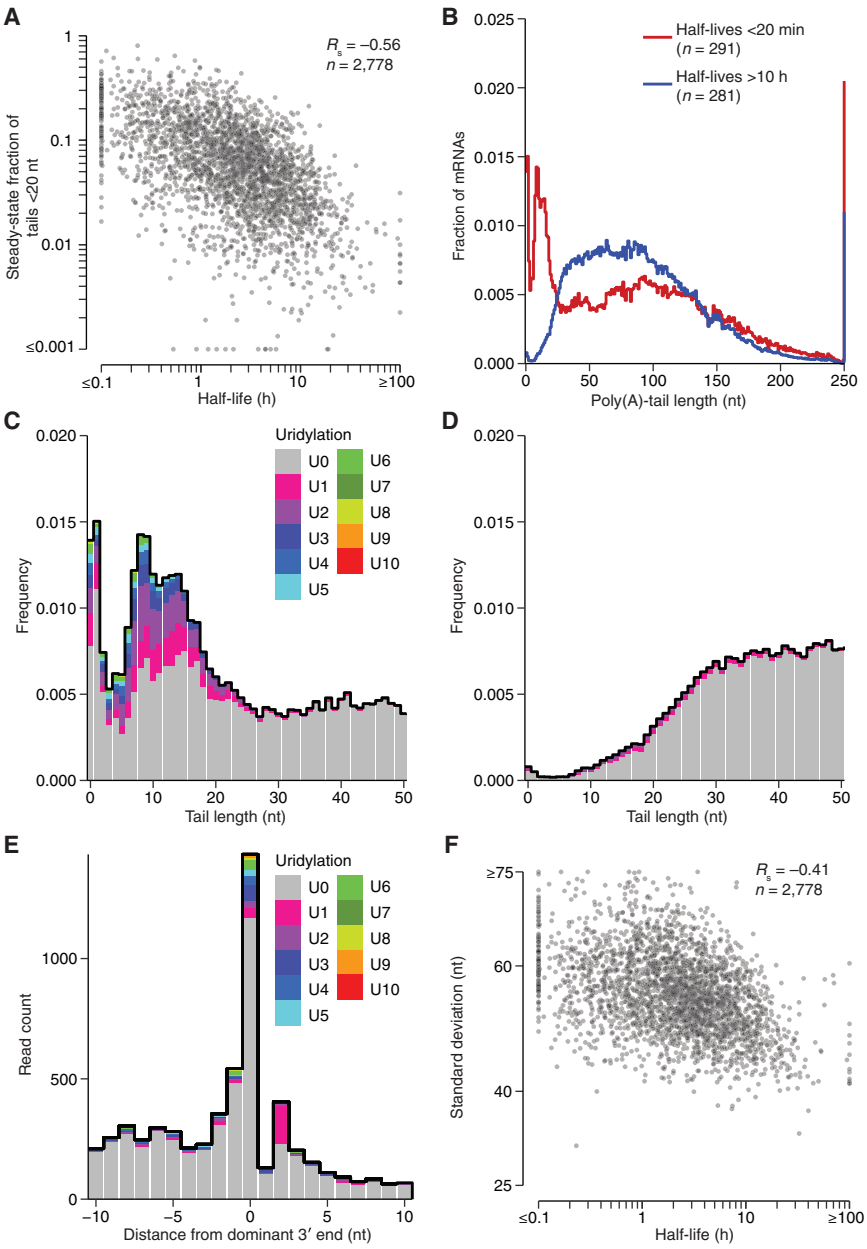


Figure 6. A Modest Buildup of Short-Tailed Isoforms of Short-Lived mRNAs

(A) Relationship between the steady-state fraction of tails < 20 nt and mRNA half-life. For mRNAs of each gene, the fraction of tails < 20 nt was calculated from a composite distribution generated as in Figure 2A, which accounted for very short and highly modified tails.

(B) Metatranscript distributions of steady-state tail lengths of short- and long-lived mRNAs (red and blue, respectively), with mRNAs from each gene contributing density according to their abundance. Results were almost identical when mRNAs were weighted such that each gene contributed equally. This analysis used the composite distributions as in (A).

(C) Uridylation of short-lived mRNAs with short poly(A) tails. For mRNAs with half-lives < 20 min, the fraction of molecules with the indicated poly(A)-tail length at steady state is plotted, indicating for each tail length the proportion of tails appended with 0 through 10 U nucleotides (key). For mRNAs with poly(A)-tail length of 0, U residues were counted only if they could not have been genomically encoded. As poly(A) tails approached 20 nt, the ability to map reads with ≥ 3 terminal U residues diminished, but the ability to map reads with 1–2 terminal U residues was retained for poly(A) tails of each length.

(D) Uridylation of long-lived mRNAs (half-lives > 10 h) with short poly(A) tails. Otherwise as in (C).

(E) Distribution of tailless tags (regardless of mRNA half-life) as a function of their distance from the annotated 3′ end of the UTR. Tags with a terminal A (or with a terminal A followed by one or more untemplated U) were excluded, even if the A might have been genomically encoded. The proportion of tails appended with 0 through 10 U nucleotides is shown (key).

(F) Relationship between the standard deviation of steady-state tail length and mRNA half-life. Otherwise as in (A).

See also Figures S5C–S5J.

tails of < 20 nt (Figure 6B). Closer inspection of these two peaks revealed that these short-tailed isoforms of short-lived mRNAs were dramatically enriched in mono- and oligouridylated termini (Figures 6C, 6D, and S5D), consistent with studies showing that uridylation occurs preferentially on shorter tails and helps to destabilize mRNAs (Kwak and Wickens, 2007; Rissland et al., 2007; Rissland and Norbury, 2009; Chang et al., 2014; Lim et al., 2014), and further indicating that uridylation occurs preferentially on short-lived mRNAs.

The observation of a 0–1-nt peak in the steady-state tail-length distribution prompted examination of fully deadenylated isoforms of mRNAs that were initially polyadenylated. Molecules without tails were often also missing the last few nucleotides of the 3′ UTR (Figures 6E and S5E), suggesting that after removing the tail, the deadenylation machinery (or some other 3′-to-5′ exonuclease) usually proceeds several nucleotides into the mRNA body. Analysis of mRNAs with tails indicated that, with few exceptions, the last nucleotide of the 3′ UTR was consistently defined (Figures S5F–S5H), which supported the idea that the missing nucleotides of tailless molecules had not been lost during the process of cleavage and polyadenylation. Analysis of the final dinucleotides of tailless tags revealed no consistent pattern after accounting for the genomic background, suggesting that other factors, such as proteins or more distal nucleotide composition, influence the position at which the exonuclease stops.

Despite their presence, the two peaks of short-tailed isoforms did not dominate the distribution, as most short-lived mRNAs (70%) had tails exceeding 30 nt (Figure 6B). Indeed, compared to long-lived mRNAs, these short-lived mRNAs also had modest enrichment for very long tails (> 175 nt) (Figures 6B, S5I, and S5J), perhaps due to an initial lag in assembling deadenylation machinery as mRNAs enter the cytoplasm, which would cause a relatively larger fraction of short-lived mRNAs to exist in the cytoplasm prior to an initial encounter with a deadenylase. The increased fractions of both short-tail and long-tail isoforms for short-lived mRNAs led to broader overall tail-length distributions (Figure 6B) with increased standard deviations in tail length (Figure 6F; $R_s = -0.41$). Moreover, the increased fractions of shorter and longer isoforms offset each other when calculating mean tail length, leading to similar mean tail lengths for the short- and long-lived mRNAs (Figure S5K; median mean tail lengths = 89 and 92 nt, respectively), which contributed to the lack of correlation between half-life and mean tail length at steady state (Figure 2A). Most importantly, the low magnitude of the buildup supported our conclusion that for most mRNAs the steps of deadenylation and subsequent decay are kinetically coupled: short-tailed mRNAs that had previously undergone more rapid deadenylation are more rapidly degraded. This coupling prevents a large buildup of short-tailed isoforms of rapidly deadenylated RNAs, thereby enabling the large range in deadenylation rate constants to impart a similarly large range in mRNA stabilities.

Deadenylation and Decay Dynamics of Synchronous mRNA Populations

Our continuous-labeling experiments were designed to measure the dynamics of mRNA metabolism in an unperturbed cellular environment. However, this framework required deadenylation and deadenylation-dependent decay parameters to be inferred as mRNAs from each gene approached their steady-state expression levels and tail lengths, with their populations becoming progressively less synchronous, causing the signal for their end behavior to be diluted. For orthogonal measurements of these parameters, we performed a pulse-chase-like experiment that more closely resembled previous studies with single-gene reporters, in that it monitored synchronous populations of mRNAs from each gene. After a 1-h pulse of 5EU, 3T3 cells were treated with actinomycin D (actD) to block transcription, and abundances and poly(A)-tail lengths of the mRNAs produced during the 5EU-labeling period were measured over the next 15 h, thereby revealing the behavior of synchronized mRNA populations as they age (Figure 7A).

As expected, tail lengths of labeled mRNAs progressively decreased after transcriptional inhibition, with median lengths shortening from 123 to 51 nt over the course of the experiment (Figure 7B). Examination of mean tail lengths of mRNAs from each gene revealed a similar trend (Figure 7C). At later time points mean tail-length distributions peaked between 45 and 50 nt (Figure 7C), far below the 100–105-nt mode of the steady-state distribution, which included mRNAs of all ages (Figure 1C).

The actD treatment had some side effects. At later time points, a ~30-nt periodicity emerged in the single-molecule tail-length distributions (Figure 7B). Although such phasing of tail lengths, with a period resembling the size of a PABPC footprint, has been observed in mammalian cells following CCR4 knockdown (Yi et al., 2018) and in *C. elegans* (Lima et al., 2017), only subtle phasing was observed in unperturbed mammalian cells (Figure 6B). The more prominent periodicity observed after prolonged actD treatment was presumably the result of more dense packing of PABPC on poly(A) tails in the context of a diminishing mRNA pool. A second side effect of actD treatment concerned mRNA half-lives, which increased from a median of 2.1 h in the continuous-labeling experiment to a median of 3.8 h in the transcriptional-shutoff experiment (Figure S3E). This increase was observed even for mRNAs with the shortest half-lives, which indicated that it occurred before actD could have influenced protein output, i.e., in less time than that required for mRNA nucleocytoplasmic export and translation. This result generalized previous observations concerning the effects of actD on reporter-mRNA stabilities (Chen et al., 1995).

Despite the side effects of actD, the rank order of mRNA half-lives determined from the transcriptional-shutoff experiment agreed well with that from the continuous-labeling experiment (Figure S3E; $R_s = 0.78$), indicating that the transcriptional-shutoff experiment captured key aspects of the unperturbed condition. In addition, mRNA half-lives calculated from the continuous-labeling experiment strongly corresponded to mean tail length observed 1 h after actD treatment (Figure 7D; note that 1 h after actD treatment was 2 h after 5EU labeling and thus most comparable to Figure 2B). Indeed, the strength of the correspondence between half-life and 1-h tail length ($R_s = 0.88$) further supported our conclusion that the vast majority of mRNAs are primarily degraded through deadenylation-linked mechanisms.

To further analyze the results of the transcriptional-shutoff experiment, we grouped mRNAs into cohorts based on their half-lives and monitored the abundance and average tail length of mRNAs from individual genes at each time point (Figure 7E). Regardless of mRNA half-life, tails initially shortened with little change in abundance until mean tail lengths fell below 100 nt. As expected based on the strong correspondence between half-life and 1-h tail length (Figure 7D), mRNAs with shorter half-lives underwent more rapid tail shortening (Figure 7E). Once mean tail lengths fell below 50 nt (implying that a substantial fraction of tails fell below 25 nt), degradation accelerated. This acceleration was more prominent for mRNAs with shorter half-lives, which confirmed our conclusion that short-tailed mRNAs that had undergone more rapid deadenylation are also more rapidly degraded (Figure 7E).

To examine how well our model predicted this behavior, we used it to predict the results of the transcriptional-shutoff experiment, using the rate constants measured earlier from the continuous-labeling experiment. When simulating a shorter time course to account for the more rapid deadenylation and decay observed without actD, the results predicted by the model agreed well with the experimental observations ($R_s = 0.93$ and 0.61 for mean tail length and abundance, respectively; $n = 11{,}273$ values above the abundance threshold for 2,687 mRNAs), including the precipitous decline in abundance when mean tail lengths fell below 50 nt and the faster degradation of short-tailed mRNAs that had undergone faster deadenylation (Figure 7F). The striking correspondence between the predictions of the model, which had been trained on the continuous-labeling experiment, and the observations of the transcriptional-shutoff experiment validated the

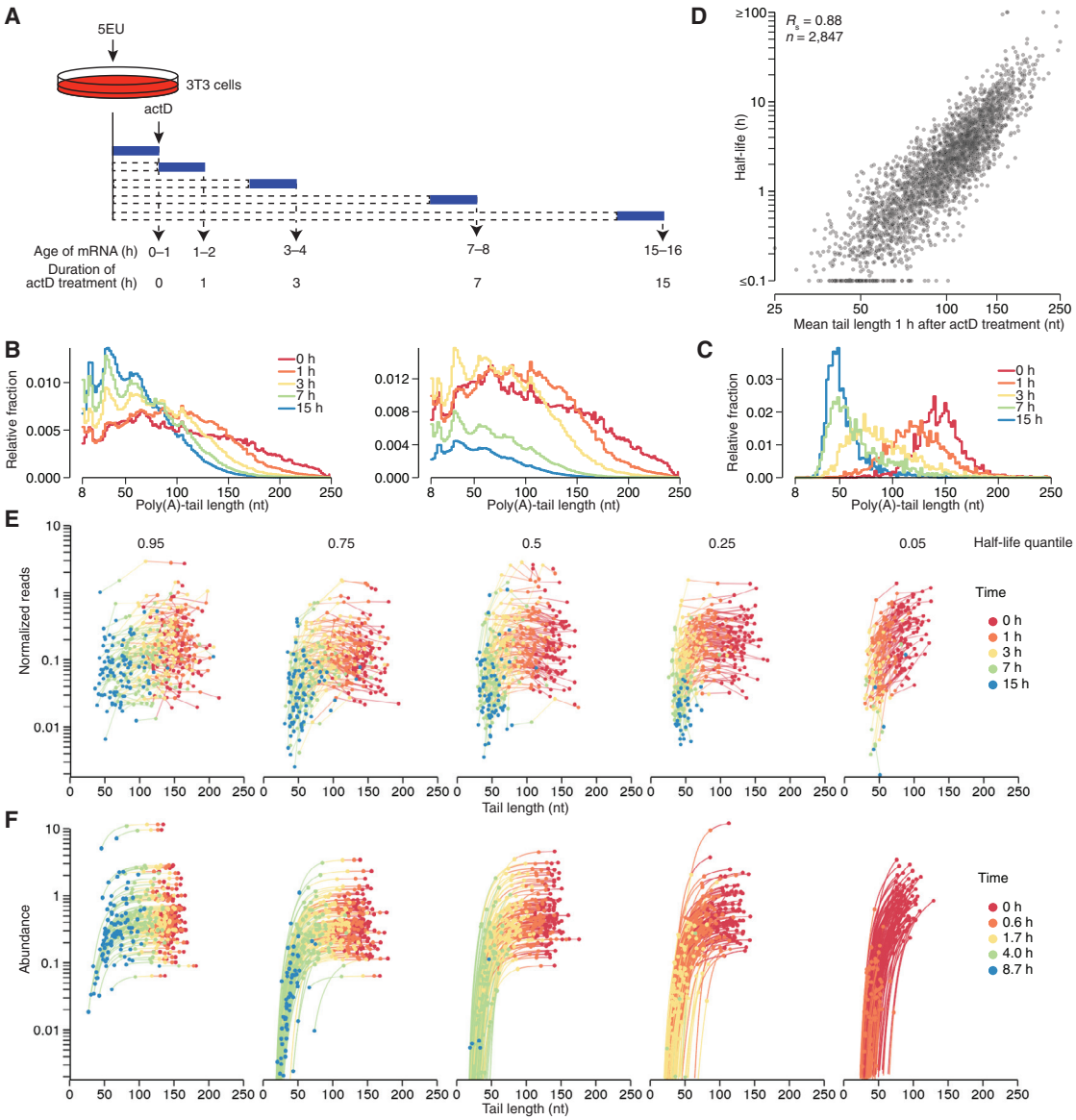


Figure 7. Deadenylation and Decay Dynamics of Synchronous mRNA Populations

(A) Schematic of 5EU metabolic-labeling and actD treatments used to analyze synchronized cellular mRNAs. Cells from cell line 2 were treated for 1 h with 5EU, then treated with actD continuously over a time course spanning 15 h.

(B) Tail-length distributions of labeled mRNA molecules observed at the indicated times after stopping transcription (key). Left: distributions were normalized to all have the same area. Right: distributions were scaled to the abundance of labeled RNAs in each sample and then normalized such that the 0-h time interval had an area of 1. Each bin is 2 nt; results for the bins with tail lengths < 8 nt and ≥ 250 nt are not shown. At 0 h, 7% of the tails were still ≥ 250 nt, which helps explain why the density for the remainder of the tails fell below that observed at 1 h.

(C) Distributions of mean poly(A)-tail lengths for labeled mRNAs of each gene after the indicated duration of transcriptional shutoff. Values for all mRNAs that passed the cutoffs for tail-length measurement at all time points were included (n = 2,155). Each bin is 2 nt.

(D) Relationship between half-life and mean tail length of labeled mRNAs from each gene after 1 h of actD treatment.

(E) Labeled mRNA abundance as a function of mean tail length over time. Results are shown for mRNAs grouped by half-life quantiles (95%, 75%, 50%, 25%, and 5%, left to right, with mRNAs in the 5% bin having the shortest half-lives). Each half-life bin contains 100 genes. mRNA abundance was determined from paired RNA-seq data. Each line connects values for mRNA from a single gene.

(F) Simulation of mRNA abundance as a function of mean tail length over time. For each gene in (E), model parameters fit from the continuous-labeling experiment were used to simulate the initial production of mRNA and its mean tail length from each gene, as well as the fates of these mRNAs and mean tail lengths after production rates were set to 0. Results are plotted as in (E), but using a shorter time course (key) to accommodate the faster dynamics observed without actD. See also Figure S3E.

results and conclusions from both experiments as well as from our analytical framework.

DISCUSSION

Previous studies provide information on deadenylation and degradation dynamics for four mammalian mRNAs and some derivatives, with deadenylation rates reported for two of these four (Mercer and Wake, 1985; Wilson and Treisman, 1988; Shyu et al., 1991; Chen et al., 1995; Gowrishankar et al., 2005; Yamashita et al., 2005). Our study provided a more comprehensive resource for deriving the principles of cytoplasmic mRNA metabolism. Initial analyses revealed unanticipated intra- and intergenic variability in initial tail lengths and indicated that almost all endogenous mRNAs are degraded primarily through deadenylation-linked mechanisms, implying that the deadenylation rate of each mRNA largely determines its half-life with surprisingly little contribution from other mechanisms, such as endonucleolytic cleavage and deadenylation-independent decapping.

Mathematical modeling of our data expanded the known range in deadenylation rate constants from 60-fold to 1000-fold and showed that the link between deadenylation rate and decay generally operates at two levels. First, mRNAs with faster deadenylation rate constants more rapidly reach the short tail lengths associated with destruction of the mRNA body. With respect to the reason that short tail lengths trigger decay, our analyses support the prevailing view that loss of PABPC binding to the poly(A) tail enhances decay, with destabilization beginning as tails become too short for cooperative binding of a PABPC dimer and accelerating as tails become too short for efficient binding of a single PABPC molecule.

A more rapid approach to short-tailed isoforms is not the whole story. mRNAs with identical 20-nt tails but from different genes can have widely different decay rate constants (1000-fold). Moreover, there is a logic to these differences—a logic conferred by the second link between deadenylation rate and decay: mRNAs that had previously undergone more rapid deadenylation decay more rapidly upon reaching short tail lengths. The coherent regulation of deadenylation and short-tailed mRNA decay rates functionally integrates mRNA turnover into a single process to ensure that mRNAs that are rapidly deadenylated are also rapidly cleared from the cell. With respect to mechanism, perhaps changes that occur as mRNA-protein complexes are remodeled to enhance deadenylation also recruit the decapping machinery and its coactivators. Terminal uridylation, which is known to stimulate decapping (Rissland and Norbury, 2009; Morozov et al., 2010; Lim et al., 2014), may aid in this remodeling, as uridylation was preferentially observed on rapidly deadenylated, short-lived mRNAs. Physical connections between the CCR4-NOT deadenylase complex and the decapping complex (Haas et al., 2010; Ozgur et al., 2010; Jonas and Izaurralde, 2015), as well as the intracellular colocalization of these complexes (Parker and Sheth, 2007), presumably also help coordinate deadenylation and short-tailed mRNA decay rates.

The large differences observed for both deadenylation and deadenylation-dependent decay rate constants of mRNAs from different genes raise the question of what mRNA features might specify these differences. MicroRNAs and other factors that help recruit deadenylase complexes typically bind to sites in 3′ UTRs, implying that these sites help to specify the differences (Mauxion et al., 2009; Mühlemann and Lykke-Andersen, 2010; Vlasova-St Louis and Bohjanen, 2011; Van Etten et al., 2012; Fabian et al., 2013; Leppek et al., 2013; Du et al., 2016; Bartel, 2018). However, global analyses of tandem UTR isoforms indicate that the magnitude of the differences conferred by 3′-UTR sequences in NIH 3T3 cells is relatively modest (Spies et al., 2013). Codon composition can also contribute to differences in mRNA stability, but this contribution explains only a small fraction of the variability observed for endogenous mRNAs of mammalian cells (Presnyak et al., 2015; Radhakrishnan et al., 2016; Forrest et al., 2018; Wu et al., 2019). Additional insight will be required to account more fully for the large differences in stabilities observed for different mRNAs. Our results indicate that the focus should be on sequences and processes that influence or correlate with deadenylation rates.

Our global observation that mRNAs typically degrade only after their tail lengths shorten extended to the mammalian transcriptome the notion that exponential decay is not fully appropriate for modeling mRNA degradation (Shyu et al., 1991; Cao and Parker, 2001; Trcek et al., 2011; Deneke et al., 2013). For the exponential model to be appropriate, an mRNA would need to have the same probability of decaying at any point after entering the cytoplasm. In contrast, recently exported, long-tailed mRNAs typically underwent little if any decay, which supported the restricted-degradation model in which mRNAs are provided a discrete time window to function in the cytoplasm. During this window, the body of the mRNA is unaltered, but its age and lifespan are tracked and determined through the action of tail-length dynamics. Nonetheless, for some analyses we used the exponential model and referred to its decay parameter as "half-life" when fitting abundance changes over time because in those cases a more complex model did not provide additional insight, and using mRNA half-lives is still common practice in the field.

Despite the utility of our mathematical model, it did not capture some finer details of mRNA metabolism. For example, it was not designed to model the burst of deadenylation that typically accompanies the loss of each terminal PABPC molecule (Webster et al., 2018). However, when considering the aggregate behavior of multiple mRNAs from the same gene, these bursts become blurred, with some molecules in the burst phase and others between bursts. Accordingly, we fit a single, continuous deadenylation rate constant for the mRNAs of each gene. Likewise, we fit a single, continuous production rate constant for the mRNAs of each gene, despite the known burst behavior of transcription initiation when examined in single cells (Cai et al., 2008).

The uniform deadenylation rate constants of the model were also not suitable for capturing aspects of tail behavior that occurred as tails fell below 20 nt. For example, our analysis of steady-state data revealed buildups of isoforms of short-lived mRNAs at two tail-length ranges: 0–1 and 7–15 nt. A model with uniform deadenylation rate constants can potentially explain a peak at 0 nt but not one at an intermediate tail length, such as 7–15 nt. Recognizing this limitation but still wanting to accurately account for the buildup of isoforms with tails < 20 nt observed for short-lived mRNAs, we fit the abundance of tails < 20 nt by averaging abundance over this length range and

comparing this average to that predicted by the model—an approach that did not require additional parameters to model a buildup of 7–15-nt tails. Such parameters might be warranted if further study shows that the fate of mRNAs with 7–15-nt tails differs from that of mRNAs with 0-nt tails—studies that can be contemplated now that the existence of this buildup is known. Another aspect of mRNA metabolism remaining to be incorporated into a mathematical model is terminal uridylation, which was particularly prominent on short-tailed isoforms of short-lived mRNAs.

A recent study observed that cytoplasmic noncanonical poly(A) polymerases can extend tails, acting on longer-tailed mRNAs and adding mostly A residues but also sometimes generating a mixed tail including a G or another non-A nucleotide (Lim et al., 2018). Because most mRNAs with these mixed tails would not be detected by PAL-seq, these mRNAs would have appeared to have been degraded in our analysis. Thus, our observation of little-to-no degradation of long-tailed mRNAs indicated that, in 3T3 cells, mRNAs with mixed tails comprised only a small fraction of the mRNA molecules at any point in time and did not impact the overall conclusions of our study.

Although our current approach does not model all aspects of mRNA metabolism, there is every reason to believe that the broad behaviors observed in these initial analyses will continue to be observed in more detailed representations of mRNA metabolism. With the acquisition of suitable pre-steady-state data, the dynamics of tail-length changes in the 0–20-nt range, of terminal uridylation, and of cytoplasmic polyadenylation could be better characterized—ultimately enabling incorporation of these phenomena into a comprehensive model of mRNA metabolism. Our methods and analytical framework offer inspiration as well as a foundation for these future efforts.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell Lines and Cell Culture
- METHOD DETAILS
  - Metabolic-Labeling Time Courses
  - RNA Standards
  - Biotinylation of 5EU Labeled RNA
  - Purification of Biotinylated RNA
  - PAL-Seq v2
  - PAL-Seq v2 Data Analysis
  - Analysis of PAL-Seq ss-Ligation Data
  - TAIL-Seq
  - RNA-Seq
  - Calculation of mRNA Half-Lives
  - Model of mRNA Metabolism
  - Bootstrap Analysis
  - Background Subtraction for PAL-Seq Data
  - ActD Treatment
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2019.12.005.

ACKNOWLEDGMENTS

We thank J. Kwasnieski and other members of the Bartel lab for helpful discussions and the Whitehead Genome Technology Core for high-throughput sequencing. This research was supported by NIH grants GM061835 and GM118135 (D.P.B.) and an NSF Graduate Research Fellowship (T.J.E.). A.O.S. was supported by NIH Medical Scientist Training Program fellowship T32GM007753. D.P.B. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

A.O.S., S.W.E., T.J.E., and D.P.B. conceived the project and designed the study. T.J.E., S.W.E., and A.O.S. performed the molecular experiments and analysis. T.J.E. performed the computational modeling with input from K.S.L. and S.E.M. S.W.E., S.G., and T.J.E. adapted PAL-seq for compatibility with current Illumina technologies. K.S.L. and S.W.E. wrote the analysis pipeline for determining tail-length measurements from PAL-seq data. A.O.S., S.W.E., and T.J.E. drafted the manuscript, and T.J.E. and D.P.B. revised the manuscript with input from the other authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 10, 2019
Revised: November 25, 2019
Accepted: December 6, 2019
Published: January 2, 2020

REFERENCES

Baer, B.W., and Kornberg, R.D. (1983). The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein. J. Cell Biol. *96*, 717–721.

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell *173*, 20–51.

Bönisch, C., Temme, C., Moritz, B., and Wahle, E. (2007). Degradation of hsp70 and other mRNAs in Drosophila via the 5′ 3′ pathway and its regulation by heat shock. J. Biol. Chem. *282*, 21818–21828.

Cai, L., Dalal, C.K., and Elowitz, M.B. (2008). Frequency-modulated nuclear localization bursts coordinate gene regulation. Nature *455*, 485–490.

Cao, D., and Parker, R. (2001). Computational modeling of eukaryotic mRNA turnover. RNA *7*, 1192–1212.

Carballo, E., Lai, W.S., and Blackshear, P.J. (2000). Evidence that tristetraprolin is a physiological regulator of granulocyte-macrophage colony-stimulating factor messenger RNA deadenylation and stability. Blood *95*, 1891–1899.

Chang, H., Lim, J., Ha, M., and Kim, V.N. (2014). TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications. Mol. Cell *53*, 1044–1052.

Chen, C.Y., and Shyu, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem. Sci. *20*, 465–470.

Chen, C.Y., Xu, N., and Shyu, A.B. (1995). mRNA decay mediated by two distinct AU-rich elements from c-fos and granulocyte-macrophage colony-stimulating factor transcripts: different deadenylation kinetics and uncoupling from translation. Mol. Cell. Biol. *15*, 5777–5788.

Dahleh, M., Dahleh, M.A., and Verghese, G. (2004). Lectures on dynamic systems and control. A+ A *4*, 1–100. https://ocw.mit.edu/courses/electrical-

engineering-and-computer-science/6-241j-dynamic-systems-and-control-spring-2011/readings/.

Decker, C.J., and Parker, R. (1993). A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. Genes Dev. *7*, 1632–1643.

Dellavalle, R.P., Petersen, R., and Lindquist, S. (1994). Preferential deadenylation of Hsp70 mRNA plays a key role in regulating Hsp70 expression in Drosophila melanogaster. Mol. Cell. Biol. *14*, 3646–3659.

Deneke, C., Lipowsky, R., and Valleriani, A. (2013). Complex degradation processes lead to non-exponential decay patterns and age-dependent decay rates of messenger RNA. PLoS ONE *8*, e55442.

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

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C.C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U.H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA *14*, 1959–1972.

Du, H., Zhao, Y., He, J., Zhang, Y., Xi, H., Liu, M., Ma, J., and Wu, L. (2016). YTHDF2 destabilizes m(6)A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. Nat. Commun. *7*, 12626.

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

Eisen, T.J., Eichhorn, S.W., Subtelny, A.O., and Bartel, D.P. (2020). MicroRNAs cause accelerated decay of short-tailed target mRNAs. Molecular Cell. https://doi.org/10.1016/j.molcel.2019.12.004.

Fabian, M.R., Cieplak, M.K., Frank, F., Morita, M., Green, J., Srikumar, T., Nagar, B., Yamamoto, T., Raught, B., Duchaine, T.F., and Sonenberg, N. (2011). miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT. Nat. Struct. Mol. Biol. *18*, 1211–1217.

Fabian, M.R., Frank, F., Rouya, C., Siddiqui, N., Lai, W.S., Karetnikov, A., Blackshear, P.J., Nagar, B., and Sonenberg, N. (2013). Structural basis for the recruitment of the human CCR4-NOT deadenylase complex by tristetraprolin. Nat. Struct. Mol. Biol. *20*, 735–739.

Forrest, M.E., Narula, A., Sweet, T.J., Arango, D., Hanson, G., Ellis, J., Oberdoerffer, S., Coller, J., and Rissland, O.S. (2018). Codon usage and amino acid identity are major determinants of mRNA stability in humans. bioRxiv. https://doi.org/10.1101/488676.

Gowrishankar, G., Winzen, R., Bollig, F., Ghebremedhin, B., Redich, N., Ritter, B., Resch, K., Kracht, M., and Holtmann, H. (2005). Inhibition of mRNA deadenylation and degradation by ultraviolet light. Biol. Chem. *386*, 1287–1293.

Haas, G., Braun, J.E., Igreja, C., Tritschler, F., Nishihara, T., and Izaurralde, E. (2010). HPat provides a link between deadenylation and decapping in metazoa. J. Cell Biol. *189*, 289–302.

Hilgers, V., Teixeira, D., and Parker, R. (2006). Translation-independent inhibition of mRNA deadenylation during stress in Saccharomyces cerevisiae. RNA *12*, 1835–1845.

Hu, J., Li, Y., and Li, P. (2013). MARVELD1 inhibits nonsense-mediated RNA decay by repressing serine phosphorylation of UPF1. PLoS ONE *8*, e68291.

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Jan, C.H., Friedman, R.C., Ruby, J.G., and Bartel, D.P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature *469*, 97–101.

Jao, C.Y., and Salic, A. (2008). Exploring RNA transcription and turnover in vivo by using click chemistry. Proc. Natl. Acad. Sci. USA *105*, 15779–15784.

Jia, H., Wang, X., Liu, F., Guenther, U.P., Srinivasan, S., Anderson, J.T., and Jankowsky, E. (2011). The RNA helicase Mtr4p modulates polyadenylation in the TRAMP complex. Cell *145*, 890–901.

Jonas, S., and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. Nat. Rev. Genet. *16*, 421–433.

Kwak, J.E., and Wickens, M. (2007). A family of poly(U) polymerases. RNA *13*, 860–867.

Leppek, K., Schott, J., Reitter, S., Poetz, F., Hammond, M.C., and Stoecklin, G. (2013). Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. Cell *153*, 869–881.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lim, J., Ha, M., Chang, H., Kwon, S.C., Simanshu, D.K., Patel, D.J., and Kim, V.N. (2014). Uridylation by TUT4 and TUT7 marks mRNA for degradation. Cell *159*, 1365–1376.

Lim, J., Lee, M., Son, A., Chang, H., and Kim, V.N. (2016). mTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. Genes Dev. *30*, 1671–1682.

Lim, J., Kim, D., Lee, Y.S., Ha, M., Lee, M., Yeo, J., Chang, H., Song, J., Ahn, K., and Kim, V.N. (2018). Mixed tailing by TENT4A and TENT4B shields mRNA from rapid deadenylation. Science *361*, 701–704.

Lima, S.A., Chipman, L.B., Nicholson, A.L., Chen, Y.H., Yee, B.A., Yeo, G.W., Coller, J., and Pasquinelli, A.E. (2017). Short poly(A) tails are a conserved feature of highly expressed genes. Nat. Struct. Mol. Biol. *24*, 1057–1063.

Mauxion, F., Chen, C.Y., Séraphin, B., and Shyu, A.B. (2009). BTG/TOB factors impact deadenylases. Trends Biochem. Sci. *34*, 640–647.

Mercer, J.F., and Wake, S.A. (1985). An analysis of the rate of metallothionein mRNA poly(A)-shortening using RNA blot hybridization. Nucleic Acids Res. *13*, 7929–7943.

Mor, A., Suliman, S., Ben-Yishay, R., Yunger, S., Brody, Y., and Shav-Tal, Y. (2010). Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. Nat. Cell Biol. *12*, 543–552.

Morozov, I.Y., Jones, M.G., Razak, A.A., Rigden, D.J., and Caddick, M.X. (2010). CUCU modification of mRNA promotes decapping and transcript degradation in Aspergillus nidulans. Mol. Cell. Biol. *30*, 460–469.

Mühlemann, O., and Lykke-Andersen, J. (2010). How and where are nonsense mRNAs degraded in mammalian cells? RNA Biol. *7*, 28–32.

Muhlrad, D., Decker, C.J., and Parker, R. (1994). Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5′–>3′ digestion of the transcript. Genes Dev. *8*, 855–866.

Nelson, J.O., Moore, K.A., Chapin, A., Hollien, J., and Metzstein, M.M. (2016). Degradation of Gadd45 mRNA by nonsense-mediated decay is essential for viability. eLife *5*, e12876.

Nocedal, J. (1980). Updating Quasi-Newton Matrices with Limited Storage. Math Comput *35*, 773–782.

Oliphant, T.E. (2007). Python for scientific computing. Comput. Sci. Eng. *9*, 10–20.

Ozgur, S., Chekulaeva, M., and Stoecklin, G. (2010). Human Pat1b connects deadenylation with mRNA decapping and controls the assembly of processing bodies. Mol. Cell. Biol. *30*, 4308–4323.

Palatnik, C.M., Storti, R.V., and Jacobson, A. (1979). Fractionation and functional analysis of newly synthesized and decaying messenger RNAs from vegetative cells of Dictyostelium discoideum. J. Mol. Biol. *128*, 371–395.

Park, E., and Maquat, L.E. (2013). Staufen-mediated mRNA decay. Wiley Interdiscip. Rev. RNA *4*, 423–435.

Parker, R., and Sheth, U. (2007). P bodies and the control of mRNA translation and degradation. Mol. Cell *25*, 635–646.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., and Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111–1124.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.

Radhakrishnan, A., Chen, Y.H., Martin, S., Alhusaini, N., Green, R., and Coller, J. (2016). The DEAD-Box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. Cell *167*, 122–132.e9.

R Core Team (2019). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Rissland, O.S., and Norbury, C.J. (2009). Decapping is preceded by 3′ uridylation in a novel pathway of bulk mRNA turnover. Nat. Struct. Mol. Biol. *16*, 616–623.

Rissland, O.S., Mikulasova, A., and Norbury, C.J. (2007). Efficient RNA polyuridylation by noncanonical poly(A) polymerases. Mol. Cell. Biol. *27*, 3612–3624.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Shav-Tal, Y., Darzacq, X., Shenoy, S.M., Fusco, D., Janicki, S.M., Spector, D.L., and Singer, R.H. (2004). Dynamics of single mRNPs in nuclei of living cells. Science *304*, 1797–1800.

Sheiness, D., and Darnell, J.E. (1973). Polyadenylic acid segment in mRNA becomes shorter with age. Nat. New Biol. *241*, 265–268.

Shyu, A.B., Belasco, J.G., and Greenberg, M.E. (1991). Two distinct destabilizing elements in the c-fos message trigger deadenylation as a first step in rapid mRNA decay. Genes Dev. *5*, 221–231.

Soetaert, K., Petzoldt, T., and Setzer, R.W. (2010). Solving Differential Equations in R. R J *2*, 5-15.

Spies, N., Burge, C.B., and Bartel, D.P. (2013). 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res. *23*, 2078–2090.

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

Tani, H., Imamachi, N., Salam, K.A., Mizutani, R., Ijiri, K., Irie, T., Yada, T., Suzuki, Y., and Akimitsu, N. (2012). Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability. RNA Biol. *9*, 1370–1379.

Trcek, T., Larson, D.R., Moldón, A., Query, C.C., and Singer, R.H. (2011). Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. Cell *147*, 1484–1497.

Tullai, J.W., Schaffer, M.E., Mullenbrock, S., Sholder, G., Kasif, S., and Cooper, G.M. (2007). Immediate-early and delayed primary response genes are distinct in function and genomic architecture. J. Biol. Chem. *282*, 23981–23995.

Van Etten, J., Schagat, T.L., Hrit, J., Weidmann, C.A., Brumbaugh, J., Coon, J.J., and Goldstrohm, A.C. (2012). Human Pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. J. Biol. Chem. *287*, 36370–36383.

Vlasova-St Louis, I., and Bohjanen, P.R. (2011). Coordinate regulation of mRNA decay networks by GU-rich elements and CELF1. Curr. Opin. Genet. Dev. *21*, 444–451.

Webster, M.W., Chen, Y.H., Stowell, J.A.W., Alhusaini, N., Sweet, T., Graveley, B.R., Coller, J., and Passmore, L.A. (2018). mRNA deadenylation is coupled to translation rates by the differential activities of Ccr4-Not nucleases. Mol. Cell *70*, 1089–1100.e8.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the Tidyverse. J. Open Source Softw. *4*, 1686.

Wilson, T., and Treisman, R. (1988). Removal of poly(A) and consequent degradation of c-fos mRNA facilitated by 3′ AU-rich sequences. Nature *336*, 396–399.

Wu, Q., Medina, S.G., Kushawah, G., DeVore, M.L., Castellano, L.A., Hand, J.M., Wright, M., and Bazzini, A.A. (2019). Translation affects mRNA stability in a codon-dependent manner in human cells. eLife *8*, e45396.

Yamashita, A., Chang, T.C., Yamashita, Y., Zhu, W., Zhong, Z., Chen, C.Y., and Shyu, A.B. (2005). Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. Nat. Struct. Mol. Biol. *12*, 1054–1063.

Yi, H., Park, J., Ha, M., Lim, J., Chang, H., and Kim, V.N. (2018). PABP cooperates with the CCR4-NOT complex to promote mRNA deadenylation and block precocious decay. Mol. Cell *70*, 1081–1088.e5.

Curriculum vitae

Sean E. McGeary

Education

2021 Ph.D., Biology
Massachusetts Institute of Technology, Cambridge, MA
2009 Sc.B. with Honors, magna cum laude, Biophysics
Brown University, Providence, RI

Research experience

2012–2021 Graduate Student, MIT
Department of Biology (Prof. David Bartel)
Studying the biochemical basis of miRNA targeting.
2009–2011 Research Technician, The Rockefeller University
Laboratory of Prof. Thomas Tuschl
Developed *in situ* hybridization diagnostic for imaging of miRNA levels in formaldehyde fixed and paraffin embedded tissues.
2008–2009 Undergraduate Research Assistant, Brown University
Department of Molecular Biology, Cell Biology and Biochemistry (Prof. Gerwald Jogl)
Crystallography of *E. coli* methyltransferase PrmA with ribosomal subunit L11.

Publications

1. **McGeary, S.E.***, Lin, K.S.*, Shi, C.Y., Pham, T.M., Bisaria, N., Kelley, G.M., and Bartel, D.P. (2019). The biochemical basis of miRNA targeting efficacy. Science *366*, eaav1741.
2. Eisen, T.J.*, Eichhorn, S.W.*, Subtelny, A.O.*, Lin, K.S., **McGeary, S.E.**, Gupta, S., and Bartel, D.P. (2019). The dynamics of cytoplasmic mRNA metabolism. Mol. Cell *77*, 786–799.
3. Denzler, R., **McGeary, S.E.**, Title, A.C., Agarwal, V., Bartel, D.P., and Stoffel, M. (2016). Impact of miRNA levels, target-site complementarity and cooperativity on ceRNA-regulated gene expression. Mol. Cell *64*, 565–579.
4. Eichhorn, S.E.*, Guo, H.*, **McGeary, S.E.**, Rodriguez-Mias, R.A., Shin, C., Baek, D., Hsu, S., Ghoshal, K., Villén, J., and Bartel, D.P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Mol. Cell *56*, 1–12.
5. Lambert, N., Robertson, A., Jangi, M., **McGeary, S.**, Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol. Cell *54*, 887–900.
6. Auyeung, V.C., Ulitsky, I., **McGeary, S.E.**, and Bartel D.P., (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell *152*, 844–858.
7. Renwick, N., Cekan, P., Masry, P.A., **McGeary, S.E.**, Miller, J.B., Hafner, M., Li, Z., Mihailovic, A., Morozov, P., Brown, M., Gogakos, T., Mobin, M.B., Snorrason, E.L., Feilotter, H.E., Zhang, X., Perlis, C.S., Wu, H., Suárez-Fariñas, M., Feng, H., Shuda, M., Moore, P.S., Tron, V.A., Chang, Y., and Tuschl, T. (2013). Multicolor microRNA FISH effectively differentiates tumor types. J. Clin. Investig. *123*, 2694–2702.

Oral presentations at international meetings

2018 High-throughput biochemical analyses of miRNA targeting reveal miRNA-specific binding preferences that markedly improve miRNA target predictions. The Complex Life of RNA; EMBL Heidelberg, Germany, October 3rd–6th
2017 Biochemical analyses of millions of possible miRNA–target site interactions, Microsymposium on small RNAs; IMP, Vienna, May 26th–28th

Honors and awards

2019 Spirit Award for exceptional contributions to the institute, Whitehead Institute
2015 Teresa Keng Graduate Teaching Prize, MIT
2012 Honorable Mention, NSF Graduate Research Fellowship Program
2007 Karen T. Romer Undergraduate Teaching & Research Award, Brown University
2006 William J. Whittle Scholarship, Brown University

Teaching experience

2015 Teaching Assistant, Introductory Biology, MIT
2014 Tutor, Principles of Biochemical Analysis, MIT
2013 Tutor, Principles of Biochemical Analysis, MIT
2012 Teaching Assistant, Principles of Biochemical Analysis, MIT