Biochemically informed modeling of miRNA targeting efficacy

by

Kathy S. Lin

B.A. Chemical and Physical Biology, Secondary Field Computer Science (2014)
Harvard University

Submitted to the Program of Computational and Systems Biology in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

February 2020

Signature of Author...
Kathy S. Lin
Computational and Systems Biology Graduate Program
October 21, 2019

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

Accepted by...
Christopher Burge
Professor of Biology
Director, Computational and Systems Biology Graduate Program

Biochemically informed modeling of miRNA targeting efficacy

by

Kathy S. Lin

Submitted to the Program of Computational and Systems Biology on
October 21, 2019 in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in
Computational and Systems Biology

Abstract

In metazoans, microRNAs (miRNAs) are short pieces of RNA that load into Argonaute (AGO) proteins and base-pair to complementary sequences in mRNAs. Upon binding an mRNA, AGO–miRNA complexes recruit machinery that translationally represses and degrades the mRNAs. Mammalian genomes encode hundreds of miRNAs, and most mRNAs in mammals have evolutionarily conserved target sites to at least one of these miRNAs. Because of the widespread and varied roles of miRNAs in regulating gene expression, there have been many efforts over the past decade to predict the extent of targeting between a miRNA and an mRNA from their sequences alone. This targeting relationship between a miRNA and an mRNA depends on the binding affinities for the AGO–miRNA complex to target sites on the mRNA, which are poorly predicted by nearest-neighbor rules used for predicting RNA–RNA duplex stabilities. This is presumably because AGO modulates the energetics of duplexes formed between its loaded miRNA and mRNA target sites.

The recent development of a high-throughput method of measuring RNA-binding affinities, RNA bind-n-seq (RBNS), has allowed us to determine the relative K_D values for AGO–miRNA complexes binding to hundreds of thousands of potential target sites. In this work, we use these biochemical parameters to build a quantitative model of miRNA targeting that predicts mRNA repression by a miRNA in cells better than existing *in silico* models. We then expand this approach to all miRNAs, including those for which we have not measured binding affinities for, by training a convolutional neural network (CNN) to predict the binding affinity between arbitrary miRNA and target sequences. We show that CNN-predicted K_D values parallel the utility of experimentally determined K_D values in predicting the repression of mRNAs in cells.

By measuring the binding affinities between miRNAs and their targets, we can also estimate how much binding affinity contributes to miRNA-mediated targeting. Although the majority of the variance in targeting is attributable to binding affinity, about 40% of the variance remains unexplained, motivating future efforts to expand the deep learning framework to learn important features of mRNAs outside of target sites that influence miRNA activity.

Thesis Advisor: David P. Bartel
Title: Professor of Biology

Acknowledgments

I have been incredibly lucky to have both amazing scientific and personal support networks throughout my graduate school career. My mentor, Dave, has taught me how to be thoughtful and rigorous in my science and precise in my writing. He generously devotes as much time as I need for advice and mentorship and has created a supportive lab environment for his trainees. I hope that, in my time in his lab, I have absorbed some amount of his work ethic, patience, and ability to think deeply about biology.

I would also like to thank my thesis advisory committee members Phil Sharp and Chris Burge. Phil always keeps me grounded in thinking about the physical interactors in biological processes. He is also an expert on RNA-binding proteins and always urges me to consider the entire cytoplasmic milieu when I become too focused on the activity of miRNAs. Chris has been a long-time collaborator of the Bartel lab, and I cannot imagine working on miRNA target prediction or RNA bind-n-seq data without his advice and input.

Chris has also advised me academically and professionally in his role as the director of the Computational and Systems Biology program at MIT, which has been a great source of support over the years. I'd especially like to thank Cassandra, Grace, Jacob, and Tristan for their friendship and for introducing me to so many different card and board games! I also want to thank Maria, Joy, and Max for their mentorship, especially when I was just starting my PhD and Jacquie, for making every milestone throughout graduate school go smoothly.

The Bartel lab has been a nurturing and intellectually stimulating environment, and I could not have asked for better lab-mates to spend graduate school with. Jamie, Stephen, Namita, Sean, Jeff, and Tim have all mentored me, gave me advice on project proposals, and helped me make challenging decisions. Vikram has been indispensable as an expert on miRNA target prediction and has continued to advise me after he left the Bartel lab. I worked especially closely with Sean on my project, as well as Charlie, Namita, Thy, and Gina, and I could not have accomplished anything in my PhD without them. The same could be said about Laura and Asia, who both work to keep the Bartel lab running smoothly. I've also had the good fortune of collaborating with Sahin from the Page lab, who is a great resource for advice on how to analyze sequencing data.

Lastly, I want to thank my family for their bottomless support and love throughout my life. My brother Milo has always been my role model, both as a scientist and as a person, and I'm excited to finally meet my new niece Terra soon! Finally, I owe a tremendous amount to my partner, Renzo, for being my rock and best friend for the past 6 years. I look forward the next chapter of our lives and whatever adventures come with it.

Table of Contents

Chapter 1. Introduction

Complexity through gene regulation

Across all known domains of life, organisms encode the necessary instructions for living and growing in their genomes. While larger genomes can contain more information and allow for more complex and potentially beneficial activities, they in turn require more physical storage space and longer copying times. Organisms have therefore evolved multilayered gene regulatory networks that combinatorially expand the space of possible gene expression states from a relatively small number of genes. The human genome, for example, only contains around 20,000 genes, and yet it can specify countless numbers of molecular, cellular, and organismal processes. Extensive regulation occurs at all levels of gene expression and can affect the production rate, degradation rate, localization, and structural conformation of each intermediate gene product.

In eukaryotes, many of the key regulatory steps occur at the messenger RNA level, as splicing largely determines the ultimate protein sequence that will be produced (Nilsen and Graveley 2010), and the levels of mature mRNAs in the cytoplasm are major determinants of protein levels (Gygi et al. 1999; Schwanhüusser et al. 2011). While steady-state mRNA levels in cells are mostly set by transcription rates (Schwanhüusser et al. 2011), degradation rates are important for determining how quickly mRNA levels respond to shifts in gene regulation, with rapidly-degraded mRNAs being more sensitive than more stable mRNAs (Yang et al. 2003). Unlike the transcription rate of mRNAs, which can be encoded in promotor or enhancer sequences (Kwak et al. 2013; Core et al. 2014), much of the information dictating the degradation rate of an mRNA needs to be written on the mRNA molecule itself so that it remains with the mRNA after exiting the nucleus.

The role of miRNAs in modulating mRNAs

For plant and animal mRNAs, one such degradation signal is recognized by a short (~22 nucleotides in length) RNA called a microRNA (miRNA) that is loaded into an Argonaute (AGO) protein to form an RNA-induced silencing complex (RISC) (Bartel 2009). In plants, the majority of the interactions between miRNAs and their target mRNAs involve Watson-Crick base-pairing along the entire length of the miRNA and result in AGO-catalyzed cleavage of the mRNA molecules (Jones-Rhoades, Bartel, and Bartel 2006). While this mode of mRNA silencing can happen in animals (Bartel 2018), the predominant way animal miRNAs lead to the decay of their targets, particularly in mammals, is by pairing with positions 2–7 of the miRNA (Doench and Sharp 2004; Lewis, Burge, and Bartel 2005), known as the “seed region,” and recruiting deadenylation and decapping proteins (Rehwinkel et al. 2005; Wu, Fan, and Belasco 2006; C. Y. A. Chen et al. 2009). In humans, there are four AGO proteins (AGO 1–4), with AGO2 generally being the most abundant and the only one capable of cleaving mRNA targets (Liu et al. 2004), although all four are capable of recruiting TNRC6, causing the downstream deadenylation of their targets (C. Y. A. Chen et al. 2009).

Because of the short pairing requirement for being targeted by a miRNA and the fact that animals can express hundreds of distinct miRNA species, virtually all mRNAs are potential targets of a miRNA. The number of possible miRNA–target interactions decreases when considering that many miRNAs and mRNAs have tissue-specific expression or are only expressed during certain developmental windows (Bartel 2018). However, it has been estimated that the majority of mammalian mRNAs harbor evolutionarily conserved target sites for endogenous miRNAs (Friedman et al. 2009) and are thus likely to be functional targets of miRNAs in the organism.

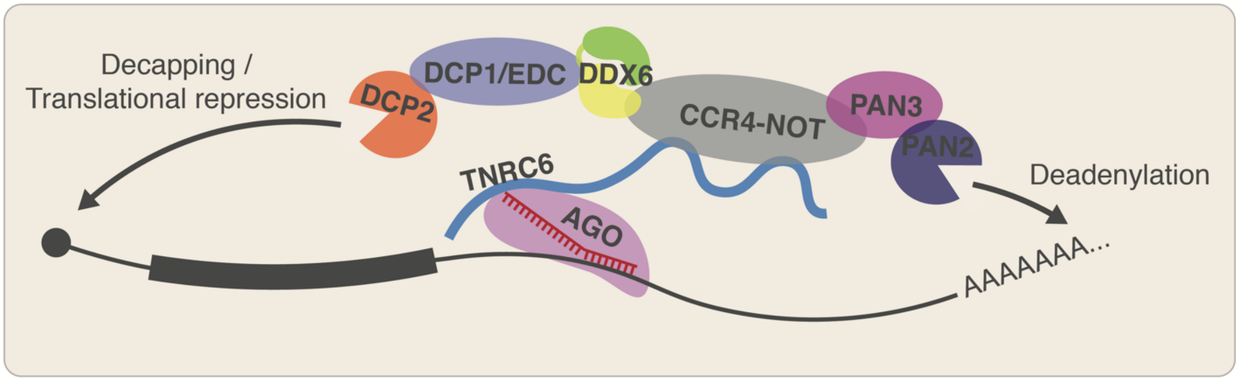


Figure 1. RISC complexes target mRNAs for translational repression and degradation. A miRNA (red) loaded into AGO (pink) recognizes target sequences complementary to positions 2–7 of the miRNA on an mRNA (black) and recruits TNRC6 (blue line), which recruits a combination of proteins that deadenylate, decap, and translationally repress the mRNA.

Upon stably binding their target mRNAs, animal miRNAs recruit degradation enzymes through the intermediate protein TNRC6 (Jonas and Izaurralde 2015) (Figure 1), which contains a long, largely unstructured domain with glycine–tryptophan (GW) repeats that can engage the tryptophan binding pockets found on AGO (Pfaff et al. 2013). TNRC6, in turn, recruits the deadenylases CCR4-NOT and the PAN2/PAN3 complex (Jonas and Izaurralde 2015) and decapping proteins DCP1 and DCP2 (Rehwinkel et al. 2005). In addition to causing the decay of mRNAs, TNRC6 also recruits DDX6, which further recruits factors that block translation initiation, causing the mRNAs to be translationally repressed as well (Jonas and Izaurralde 2015; Kamenska et al. 2016). In post-embryonic cells, this period of translational repression is shortly followed by mRNA decay, such that the mRNA decay process dominates the total effect of miRNAs on protein production (Eichhorn et al. 2014). However, during embryogenesis (at least in the context of zebrafish embryos) deadenylation of mRNAs, including by miRNAs, leads to further translational repression, rather than degradation (Subtelny et al. 2014).

Evolutionary history of miRNAs

MicroRNAs arose early in animal evolution and have been found in almost all metazoan species so far examined (Grimson et al. 2008). Over time, gene duplication events have resulted in groups of miRNAs arising from the same ancestral gene. These miRNAs often retain the same seed sequence as the ancestral gene, perhaps due to the evolutionary pressure for a functional miRNA to keep its seed sequence as changing the seed sequence by just one nucleotide can drastically alter its cohort of transcript targets. miRNAs with the same seed sequence are grouped into families, and although family members are usually paralogs, a few have independently converged to the same seed sequence (Bartel 2018). This categorization of miRNAs into families is useful because targeting specificity relies so much on the miRNA seed region.

Of the miRNA families found in humans, about 27 have been conserved since the emergence of bilaterian animals (Bartel 2018) more than 550 million years ago (Martin et al. 2000) and miRNAs as a class of RNAs have been found in as distant of a relative to humans as *Amphimedon queenslandica*, a sea sponge (Grimson et al. 2008), though no miRNAs are conserved between humans and sponges. miRNA evolution in general seems to have been quite dynamic, with no miRNAs shared between poriferans, cnidarians, and bilaterians (Grimson et al. 2008). Thus while the conserved miRNAs in humans were mostly discovered computationally (Lim et al. 2003), poorly conserved miRNAs in humans are continuously being annotated with the advent of high-throughput small-RNA sequencing techniques (Kozomara, Birgaoanu, and Griffiths-Jones 2019).

Biological roles of miRNAs

As modulators of mRNA decay rates, miRNAs have roles in maintaining and tuning mRNA levels in spatially and temporally specific ways. For example, miR-122, which constitutes the majority of the expressed miRNAs in hepatocytes, is thought to be responsible for keeping these cells terminally differentiated (Hsu et al. 2012), and deletion of miR-122 in mice leads to a number of liver diseases, including hepatitis and liver cancer (Hsu et al. 2012; Tsai et al. 2012). A number of other miRNAs have similarly crucial roles in the development and function of various tissues in model organisms, including the heart (Heidersbach et al. 2013; Wei et al. 2014), brain (Sanuki et al. 2011), immune cells (Lovat et al. 2015; Lu et al. 2010, 2015), and pancreas (Latreille et al. 2014). Perhaps the most dramatic effects of miRNAs are observed in early developmental contexts. For example, during the maternal-to-zygotic transition in zebrafish embryogenesis, miR-430 is one of the factors responsible for turning over maternally-deposited transcripts, clearing the way for the zygotic transcriptome (Giraldez et al. 2006). Knocking out both maternal and zygotic Dicer, which are critical for miRNA biogenesis, causes gastrulation and brain development defects in zebrafish during embryogenesis which can be rescued by injecting mature miR-430 into the developing embryos (Giraldez et al. 2006).

However, large-scale miRNA knockout experiments in various model organisms have revealed that while many miRNAs are crucial for normal development and function in their respective organisms, the majority of miRNAs in worms (Alvarez-Saavedra and Horvitz 2010; Miska et al. 2007), about 20% of miRNAs in flies (Y. W. Chen et al. 2014), and the majority of miRNAs in mice (C. Y. Park et al. 2012) seem to have unappreciable phenotypes when deleted. This is partly due to incomplete phenotypic studies; later studies have identified knockout phenotypes for most conserved miRNAs in mice (Bartel 2018). Some miRNAs may also appear to not have a phenotype when knocked out because they play a role in important processes that are not usually observed in lab settings. For example, miR-143 and miR-145 are co-transcribed miRNAs conserved throughout vertebrates, and yet mice lacking these two miRNAs appear to develop and function normally. A defect was only observed upon intestinal injury, after which mice lacking miR-143/145 were unable to regenerate their intestinal epithelia and died from the resulting complications whereas wild-type mice were able to recover fully in 9 days (Chivukula et al. 2014). MicroRNA families may also function redundantly with each other, which is supported by the finding that a majority of worm miRNAs surveyed do have knockout phenotypes when the knockouts are performed in sensitized backgrounds where the miRNA biogenesis pathway activity is reduced (Brenner et al. 2010).

Aside from helping to set mRNA levels in the cell, miRNAs have also been reported to play a role in reducing gene expression noise by driving higher transcription rates, primarily for lowly expressed mRNAs (Schmiedel et al. 2015). Lowly expressed mRNAs are the most impacted by stochasticity in gene expression, and miRNAs, which preferentially target lowly expressed mRNAs (Sood et al. 2006; Farh et al. 2005), may provide a mechanism for cells to ensure the steady production of these mRNAs.

Mechanisms of miRNA biogenesis and targeting

The modularity of the process of loading miRNAs into AGO makes miRNA-mediated mRNA repression versatile for both the cell and for researchers attempting to modulate gene expression. While some miRNAs are loaded better than others (Schwarz et al. 2003; Frank, Sonenberg, and Nagar 2010; Suzuki et al. 2015), AGO proteins can load any piece of RNA with any primary sequence as long as it is between 21 and 25 nucleotides in the length, contains a 5′-

monophosphate, and is paired with a complementary or nearly complementary companion sequence (known as a passenger strand or miRNA* strand) with ~2 nucleotides of 3′ overhang on each side (Bartel 2018). The passenger strand is usually the strand with the weaker base-pairing at its 3′-end (Khvorova, Reynolds, and Jayasena 2003; Schwarz et al. 2003), and is ejected upon successful loading of the miRNA (Kawamata and Tomari 2010) (Figure 2). Even though any such sequence, endogenous or otherwise, can load into AGO, pair to mRNAs, and recruit degradation machinery, most endogenous miRNAs arise from RNA hairpins within genomically-encoded transcripts and are processed via the canonical miRNA biogenesis pathway (Figure 2). This process starts in the nucleus, where the RNA molecule containing the hairpin (called a pri-miRNA) is recognized and cleaved by the enzyme DROSHA (Lee et al. 2003) to

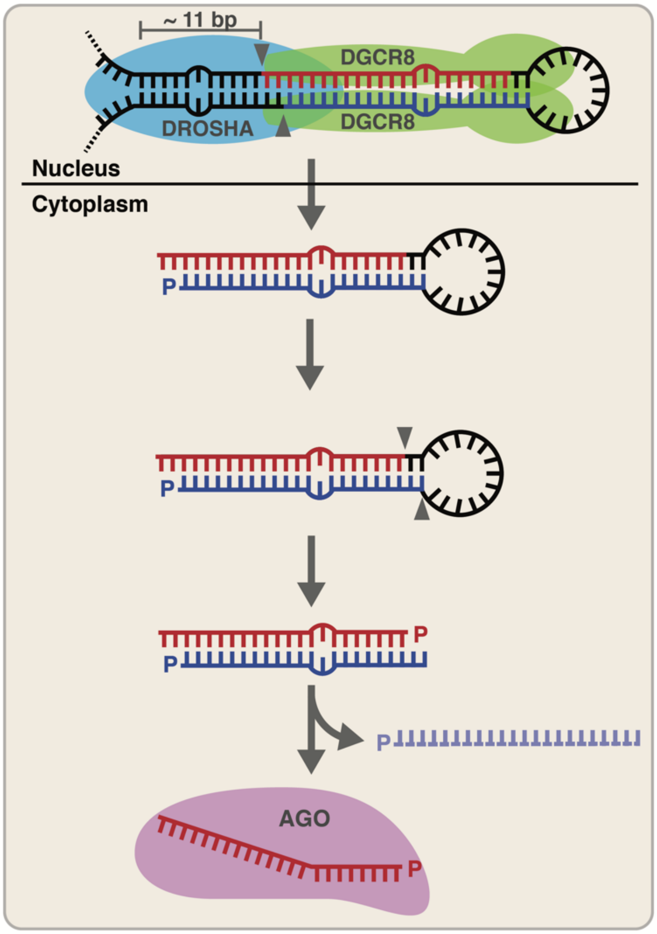


Figure 2. Canonical miRNA biogenesis pathway. After transcription, the pri-miRNA is recognized and cleaved (grey triangles) by Microprocessor in the nucleus, which consists of DROSHA (blue) and two copies of DGCR8 (green). Microprocessor measures and cleaves the pri-miRNA approximately 11 base-pairs away from the basal junction. The resulting pre-miRNA is then exported to the cytoplasm, where it is recognized and cleaved by Dicer (orange) near the single-stranded loop into a mature miRNA duplex. One strand of this duplex is loaded into AGO (pink), while the other strand is ejected and degraded.

produce a short hairpin with the first 3′ overhang. DROSHA is assisted by two copies of DGCR8, uses a combination of structural and sequence motifs to specify which RNA hairpins are eventually processed into miRNAs (Fang and Bartel 2015), and cuts them approximately 11 base-pairs away from the basal junction of the pri-miRNA stem-loop (Nguyen et al. 2015). The cleavage product, termed the pre-miRNA, is exported to the cytoplasm where it is further processed by the enzyme Dicer near the loop to produce the mature miRNA duplex (Zhang et al. 2004), which is competent to load into AGO (Figure 2).

Once a miRNA is loaded into AGO, structural work has revealed that nucleotides 2–5 of the miRNA are pre-formed into a helical structure by AGO, anticipating the conformation it would adopt upon binding its mRNA target and facilitating rapid searching of potential targets in the sea of mRNA sequences in the cytoplasm (Klum et al. 2018). Canonically, there are six types of target sites (Bartel 2009), the top four of which have been shown to robustly lead to downstream repression of their host mRNAs (Figure 3). These all involve some amount of contiguous Watson-Crick base-pairing to nucleotides 2–7 of the miRNA, with the best sites also base-pairing to position 8 of the miRNA and containing an A nucleotide opposite the first nucleotide of the miRNA (Lewis, Burge, and Bartel 2005). This preference for an A across from the 5′-most nucleotide of the miRNA, regardless of the sequence of the miRNA at this position, is conferred by a binding pocket within AGO,

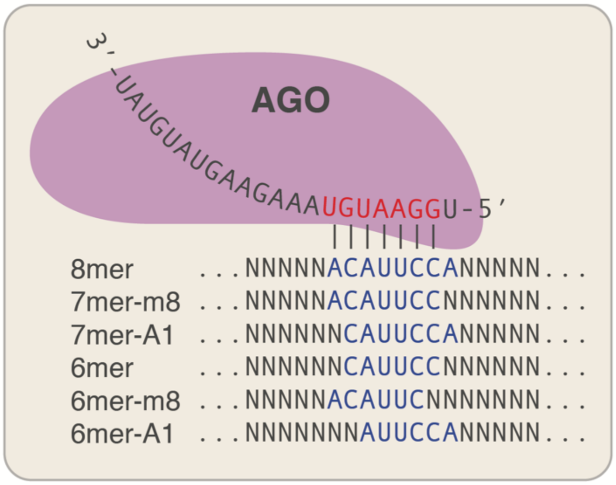

AGO
3′-UAUGUAUGAAGAAAUGUAAGGU-5′
8mer ...NNNNNACAUUCCANNNNN...
7mer-m8 ...NNNNNACAUUCCNNNNNN...
7mer-A1 ...NNNNNNCAUUCCANNNNN...
6mer ...NNNNNNCAUUCCNNNNNN...
6mer-m8 ...NNNNNACAUUCNNNNNNN...
6mer-A1 ...NNNNNNNAUUCCANNNNN...

Figure 3. Canonical site-types. miR-1 loaded into AGO (pink) base-pairing to an 8mer site embedded in an mRNA. The other 5 canonical site types are listed below the 8mer site in descending order of efficacy.

rather than base-pairing to the miRNA itself (Schirle et al. 2015). Although a small handful of functional noncanonical site types have been reported (Chi, Hannon, and Darnell 2012; Kim et al. 2016), the AGO–miRNA complex is generally intolerant of mismatches, bulges, and wobble pairings to the seed region of its miRNA (Doench and Sharp 2004), which greatly enhances the specificity a miRNA can have for its targets. Indeed, single-molecule studies have shown that while miRNAs have similar on-rates of binding to different target sequences, there is a sharp increase in off-rate when comparing miRNAs binding a "seed-matched" target and the same miRNA binding other sequences (Chandradoss et al. 2015).

In some cases, nonoptimal pairing to the seed region can be rescued by extensive complementarity (often at least five nucleotides) to the 3′-end of the miRNA (Bartel 2009). Due to their greater pairing constraints, these so-called compensatory sites are much more rare than canonical sites and harder to retain evolutionarily. However, they offer a way for mRNAs to be targeted specifically by a miRNA while avoiding cross-targeting by another miRNA with the same seed sequence. Two of the first examples discovered of a miRNA and target interaction of any kind was a pair of compensatory sites for the miRNA *let-7* in the 3′ UTR sequence of the *lin-41* mRNA in *C. elegans* (Reinhart et al. 2000; Brennecke et al. 2005). Seed-matched sites can also benefit from pairing to the 3′-end of the miRNA, and such "supplemental sites" constitute around 5% of the seed-matched sites in the human genome (Bartel 2009). In some cases, particularly when the 3′-pairing extends through to the very 3′-end of the miRNA, a target sequence can trigger the decay of the miRNA (Ameres et al. 2010). The mechanisms of this process of target RNA-directed miRNA degradation (TDMD) are currently unknown, and only a few endogenous examples have been found (Ameres et al. 2010; De La Mata and Großhans 2018; Bitetti et al. 2018; Kleaveland et al. 2018). However, the effects can be substantial, decreasing the half-life of a miRNA from days (which is typical of the average miRNA) to hours (Kingston and Bartel 2019).

Regardless of the mode of pairing, target sites to miRNAs are the most effective in the 3′ UTR of an mRNA, starting about 15 nucleotides downstream of the stop codon (Grimson et al. 2007). This is most likely due to scanning and translating ribosomes precluding AGO from binding 5′ UTR and ORF sequences and/or dislodging AGO complexes that have bound in these regions. During translation termination, this ribosome protection extends about 15 nucleotides into the 3′ UTR such that miRNA-binding sites can only avoid ribosome interference when they are 15 nucleotides downstream of the stop codon. In fact, an ineffective ORF site directly upstream from a stop codon can be converted to an effective site by moving the stop codon upstream of the site (Grimson et al. 2007), showing that the ribosome, rather than sequence context, is responsible for the reduced efficacy of sites outside the 3′ UTR. Within the 3′ UTRs of mRNAs, bioinformatic analyses have shown that sites closer to either the stop codon or the very 3′-end of the mRNA are more effective than sites in the center of the 3′ UTR (Grimson et al. 2007), although the mechanisms causing this phenomenon are currently unknown.

miRNA target prediction

Because the human genome encodes upwards of 300 different conserved miRNAs, and each one can interact with hundreds to thousands of targets, which can shift depending on the expression profiles of different cell-types, experimentally measuring each miRNA–target interaction would be a monumental feat. On top of the endogenous miRNAs, a virtually countless number of possible synthetic duplexes are competent for loading into AGO. These are mostly used by researchers to knock down transcripts through the AGO-mediated slicing pathway, and they would ideally be designed to reduce repression of off-targets. For all these reasons, there has been a long-standing interest in developing algorithms that can quantitatively predict the miRNA targeting efficacy for any arbitrary miRNA and mRNA.

One of the first successful approaches for predicting miRNA targeting involved looking for mRNAs with potential target sites in their 3′ UTRs that were complementary to positions 2–8 of a miRNA and evolutionarily conserved (Lewis et al. 2003). Because a single mismatch to the seed region of a miRNA can abrogate miRNA affinity, there was a need for a method to score the conservation level of an entire site, rather than the individual nucleotides in a site. This led to the development of the probability of conserved targeting (PCT) metric (Friedman et al. 2009), which calculates the branch-length score of a target site sequence relative to the background level of conservation of the entire 3′ UTR of the mRNA. This value is further normalized to the same metric calculated for control sequences in order to control for the differing amounts of conservation that may be conferred by dinucleotide content, rather than miRNA targeting.

Other methods have been developed to score potential target sites to a miRNA that incorporate features such as local sequence context of the site, the degree of possible supplementary pairing to the 3′-end of the miRNA, the predicted structural accessibility of the site in its 3′ UTR context, and the predicted RNA duplex stability between the miRNA and a site (Kiriakidou et al. 2004; Krek et al. 2005; Grimson et al. 2007; Garcia et al. 2011; Gumienny and Zavolan 2015; Agarwal et al. 2015). These features are regressed against some measurement of miRNA-dependent repression of the mRNA housing the site. One of the most straight-forward ways to measure miRNA activity in cells is to transfect a miRNA of interest into a cell-line that does not normally express that miRNA and measure the mRNA abundance fold-changes between transfected and mock-transfected populations either using microarrays or RNA-seq.

Because transfection datasets evaluated using microarrays are the most widely available, most miRNA target prediction models are trained on these transfection datasets and often validated on orthogonal datasets.

In addition to models that simply combine a list of features that may correlate with miRNA activity, some efforts to predict miRNA targeting efficacy attempt to construct a biochemical model of miRNA occupancy (Krek et al. 2005; Khorshid et al. 2013). Unfortunately, these models require knowledge of the binding affinities between AGO–miRNA complexes and their target mRNAs. Because only a few of these values have been determined experimentally (Wee et al. 2012; Salomon et al. 2015; Schirle, Sheu-Gruttadauria, and MacRae 2014; Schirle et al. 2015; Jo et al. 2015; Klum et al. 2018; Chandradoss et al. 2015), they must be estimated computationally. A popular method of estimating the affinity between a miRNA and a target sequence is by using nearest-neighbor (NN) rules for estimating RNA–RNA duplex stabilities (Xia et al. 1998). However, the handful of experimentally-measured binding affinities between miRNAs and their targets have shown that AGO substantially alters the energetics of binding between its loaded miRNA and a potential target (Salomon et al. 2015).

Others have attempted to use data from cross-linking and immunoprecipitation (CLIP) experiments to learn the energetics of pairing between miRNAs and their targets (Khorshid et al. 2013). In these experiments, RISC bound to RNA is cross-linked to that RNA using ultraviolet light. The resulting complexes are then immunoprecipitated, and the bound RNA is isolated and sequenced (Chi et al. 2009). These experiments provide a much more direct read-out of miRNA binding than transfection experiments because they reflect the engagement of RISC complexes on individual targets, rather than repression of entire mRNAs. However, CLIP data are subject to cross-linking biases (Lambert et al. 2014), CLIP data often contain large amounts of background binding events (Jaskiewicz et al. 2012), and it is impossible to determine fundamental binding constants from CLIP enrichments without knowledge of the concentrations of RISC in the CLIP experiments. As a result, miRNA target prediction algorithms that use CLIP data to learn miRNA–target binding affinities (Khorshid et al. 2013; Gumienny and Zavolan 2015) perform no better than the best miRNA prediction algorithms that use NN rules and site-type information (Agarwal et al. 2015) in predicting the repression of mRNAs in cells.

A high-throughput method for measuring binding affinities for RNA-binding proteins

Recently, Lambert et al. have developed a technique called RNA bind-n-seq (RBNS) for measuring the affinities of RNA-binding proteins to a random library of potential RNA binding partners *in vitro* (Lambert et al. 2014). In this technique, the RNA-binding protein of interest is tagged with an epitope that binds streptavidin, purified, and incubated with a random library of RNA molecules flanked by sequencing primers. The RNA-binding proteins are then pulled down using streptavidin beads, along with any bound RNA, the bound RNA is isolated and sequenced, and the enriched sequences are compared to those obtained by sequencing the input pool of RNA

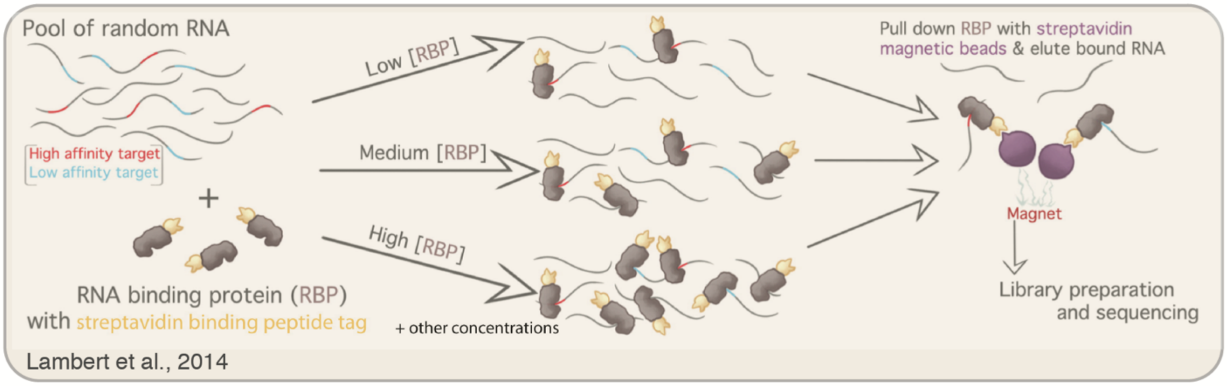

Pool of random RNA
High affinity target
Low affinity target
+
RNA binding protein (RBP)
with streptavidin binding peptide tag
Low [RBP]
Medium [RBP]
High [RBP]
+ other concentrations
Pull down RBP with streptavidin magnetic beads & elute bound RNA
Magnet
Library preparation and sequencing
Lambert et al., 2014

Figure 4. RNA bind-n-seq protocol. RNA-binding proteins tagged with a streptavidin binding tag are purified and incubated with a pool of random RNA. The RNA-binding proteins are then pulled down with streptavidin beads and their associated RNA molecules are reverse transcribed to make a cDNA library and sequenced. A range of different protein concentrations are used to capture a wide range of binding affinities. Figure adapted from Figure 1A of Lambert et al., 2014.

molecules (Figure 4). Sequence motifs that interact with the RNA-binding protein become enriched over other sequences in the pull-down library, and enrichments for individual *k*-mers can be determined (with *k* being dependent on how deeply the libraries are sequenced). This process is repeated with several different concentrations of the RNA-binding protein such that both high- and low-affinity binding interactions can be captured and quantified. This method has revealed the binding preferences for 78 human RNA-binding proteins (Dominguez et al. 2018), proving its utility and adaptability to any RNA-binding protein. RBNS is therefore an attractive method for measuring the binding affinities between RISC and its targets. The data obtained from RBNS are also theoretically sufficient for fitting the binding constants between RNA-binding proteins and *k*-mers, in addition to enrichments of *k*-mers.

Recent applications of neural networks for learning sequence binding preferences

In addition to experimental advancements that may aid miRNA target prediction, the large amounts of high-throughput sequencing data collected in recent years have fueled the development of more data-driven models of nucleic acid binding preferences. Alipahani et al. developed a deep learning model, DeepBind, that learns the binding specificities of DNA- and RNA-binding proteins and showed their model could learn from a wide range of different nucleic acid binding assays. DeepBind was also shown to outperform all other existing methods for predicting binding partners for transcription factors and RNA-binding proteins (Alipanahi et al. 2015). These types of models can pick up on hierarchically-dependent features and learn nonlinear interactions between features and can therefore be especially helpful for learning the binding specificities of RNA-binding proteins, which often have single nucleotide, di-nucleotide, and structural preferences (Y. Park and Kellis 2015). However, DeepBind and other similar models do not train on any features of the DNA- or RNA-binding proteins themselves, which means that a new model must be trained for each new DNA- or RNA-binding protein of interest. This would have to be overcome if deep learning were applied to the problem of predicting miRNA targets because each new AGO–miRNA complex is essentially a new RNA-binding protein with its own unique binding profile. Given that there are hundreds of conserved endogenous miRNAs in humans and 4^7 = 16,384 unique miRNA seed sequences possible, miRNA target prediction algorithms need to be able to generalize to arbitrary miRNA sequences without acquiring more data for each new miRNA sequence.

References

Agarwal, Vikram, George W. Bell, Jin Wu Nam, and David P. Bartel. 2015. "Predicting Effective MicroRNA Target Sites in Mammalian MRNAs." *ELife* 4 (AUGUST2015). https://doi.org/10.7554/eLife.05005.

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." *Nature Biotechnology*. https://doi.org/10.1038/nbt.3300.

Alvarez-Saavedra, Ezequiel, and H. Robert Horvitz. 2010. "Many Families of C. Elegans MicroRNAs Are Not Essential for Development or Viability." *Current Biology*. https://doi.org/10.1016/j.cub.2009.12.051.

Ameres, Stefan L., Michael D. Horwich, Jui Hung Hung, Jia Xu, Megha Ghildiyal, Zhiping Weng, and Phillip D. Zamore. 2010. "Target RNA-Directed Trimming and Tailing of Small Silencing RNAs." *Science*. https://doi.org/10.1126/science.1187058.

Bartel, David P. 2009. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell* 136 (2): 215–33. https://doi.org/10.1016/j.cell.2009.01.002.

Bartel, David P. 2018. "Metazoan MicroRNAs." *Cell*. https://doi.org/10.1016/j.cell.2018.03.006.

Bitetti, Angelo, Allison C. Mallory, Elisabetta Golini, Claudia Carrieri, Héctor Carreño Gutiérrez, Emerald Perlas, Yuvia A. Pérez-Rico, et al. 2018. "MicroRNA Degradation by a Conserved Target RNA Regulates Animal Behavior." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/s41594-018-0032-x.

Brennecke, Julius, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. 2005. "Principles of MicroRNA-Target Recognition." In *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0030085.

Brenner, John L., Kristen L. Jasiewicz, Alisha F. Fahley, Benedict J. Kemp, and Allison L. Abbott. 2010. "Loss of Individual MicroRNAs Causes Mutant Phenotypes in Sensitized Genetic Backgrounds in C. Elegans." *Current Biology*. https://doi.org/10.1016/j.cub.2010.05.062.

Chandradoss, Stanley D., Nicole T. Schirle, Malwina Szczepaniak, Ian J. Macrae, and Chirlmin Joo. 2015. "A Dynamic Search Process Underlies MicroRNA Targeting." *Cell*. https://doi.org/10.1016/j.cell.2015.06.032.

Chen, Chyi Ying A., Dinghai Zheng, Zhenfang Xia, and Ann Bin Shyu. 2009. "Ago-TNRC6 Triggers MicroRNA-Mediated Decay by Promoting Two Deadenylation Steps." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.1709.

Chen, Ya Wen, Shilin Song, Ruifen Weng, Pushpa Verma, Jan Michael Kugler, Marita Buescher, Sigrid Rouam, and Stephen M. Cohen. 2014. "Systematic Study of Drosophila MicroRNA Functions Using a Collection of Targeted Knockout Mutations." *Developmental Cell*. https://doi.org/10.1016/j.devcel.2014.11.029.

Chi, Sung Wook, Gregory J. Hannon, and Robert B. Darnell. 2012. "An Alternative Mode of MicroRNA Target Recognition." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2230.

Chi, Sung Wook, Julie B Zang, Aldo Mele, and Robert B Darnell. 2009. "Ago HITS-CLIP Decodes MiRNA-MRNA Interaction Maps." *Nature* 460 (7254): 479–86. https://doi.org/10.1038/nature08170.Ago.

Chivukula, Raghu R., Guanglu Shi, Asha Acharya, Eric W. Mills, Lauren R. Zeitels, Joselin L. Anandam, Abier A. Abdelnaby, et al. 2014. “An Essential Mesenchymal Function for MiR-143/145 in Intestinal Epithelial Regeneration.” *Cell*. https://doi.org/10.1016/j.cell.2014.03.055.

Core, Leighton J., André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. 2014. “Analysis of Nascent RNA Identifies a Unified Architecture of Initiation Regions at Mammalian Promoters and Enhancers.” *Nature Genetics*. https://doi.org/10.1038/ng.3142.

Doench, John G., and Phillip A. Sharp. 2004. “Specificity of MicroRNA Target Selection in Translational Repression.” *Genes & Development* 18 (5): 504–11. https://doi.org/10.1101/gad.1184404.

Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, et al. 2018. “Sequence, Structure, and Context Preferences of Human RNA Binding Proteins.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2018.05.001.

Eichhorn, Stephen W., Huili Guo, Sean E. McGeary, Ricard A. Rodriguez-Mias, Chanseok Shin, Daehyun Baek, Shu hao Hsu, Kalpana Ghoshal, Judit Villén, and David P. Bartel. 2014. “MRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues.” *Molecular Cell* 56 (1): 104–15. https://doi.org/10.1016/j.molcel.2014.08.028.

Fang, Wenwen, and David P. Bartel. 2015. “The Menu of Features That Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.08.015.

Farh, Kyte Kai How, Andrew Grimson, Calvin Jan, Benjamin P. Lewis, Wendy K. Johnston, Lee P. Lim, Christopher B. Burge, and David P. Bartel. 2005. “Biochemistry: The Widespread Impact of Mammalian MicroRNAs on MRNA Repression and Evolution.” *Science*. https://doi.org/10.1126/science.1121158.

Frank, Filipp, Nahum Sonenberg, and Bhushan Nagar. 2010. “Structural Basis for 5′-Nucleotide Base-Specific Recognition of Guide RNA by Human AGO2.” *Nature*. https://doi.org/10.1038/nature09039.

Friedman, Robin C., Kyle Kai How Farh, Christopher B. Burge, and David P. Bartel. 2009. “Most Mammalian MRNAs Are Conserved Targets of MicroRNAs.” *Genome Research* 19 (1): 92–105. https://doi.org/10.1101/gr.082701.108.

Garcia, David M, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. 2011. “Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of Lsy-6 and Other MicroRNAs.” *Nature Structural & Molecular Biology* 18 (10): 1139–46. https://doi.org/10.1038/nsmb.2115.

Giraldez, Antonio J., Yuichiro Mishima, Jason Rihel, Russell J. Grocock, Stijn Van Dongen, Kunio Inoue, Anton J. Enright, and Alexander F. Schier. 2006. “Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal MRNAs.” *Science*. https://doi.org/10.1126/science.1122689.

Grimson, Andrew, Kyle Kai How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. 2007. “MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.” *Molecular Cell* 27 (1): 91–105. https://doi.org/10.1016/j.molcel.2007.06.017.

Grimson, Andrew, Mansi Srivastava, Bryony Fahey, Ben J. Woodcroft, H. Rosaria Chiang, Nicole King, Bernard M. Degnan, Daniel S. Rokhsar, and David P. Bartel. 2008. "Early Origins and Evolution of MicroRNAs and Piwi-Interacting RNAs in Animals." *Nature*. https://doi.org/10.1038/nature07415.

Gumienny, Rafal, and Mihaela Zavolan. 2015. "Accurate Transcriptome-Wide Prediction of MicroRNA Targets and Small Interfering RNA off-Targets with MIRZA-G." *Nucleic Acids Research* 43 (3): 1380–91. https://doi.org/10.1093/nar/gkv050.

Gygi, Steven P., Yvan Rochon, B. Robert Franza, and Ruedi Aebersold. 1999. "Correlation between Protein and MRNA Abundance in Yeast." *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.19.3.1720.

Heidersbach, Amy, Chris Saxby, Karen Carver-Moore, Yu Huang, Yen Sin Ang, Pieter J. de Jong, Kathryn N. Ivey, and Deepak Srivastava. 2013. "MicroRNA-1 Regulates Sarcomere Formation and Suppresses Smooth Muscle Gene Expression in the Mammalian Heart." *ELife*. https://doi.org/10.7554/eLife.01323.001.

Hsu, Shu Hao, Bo Wang, Janaiah Kota, Jianhua Yu, Stefan Costinean, Huban Kutay, Lianbo Yu, et al. 2012. "Essential Metabolic, Anti-Inflammatory, and Anti-Tumorigenic Functions of MiR-122 in Liver." *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI63539.

Jaskiewicz, Lukasz, Biter Bilen, Jean Hausser, and Mihaela Zavolan. 2012. "Argonaute CLIP - A Method to Identify in Vivo Targets of MiRNAs." *Methods*. https://doi.org/10.1016/j.ymeth.2012.09.006.

Jo, Myung Hyun, Soochul Shin, Seung Ryoung Jung, Eunji Kim, Ji Joon Song, and Sungchul Hohng. 2015. "Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.04.027.

Jonas, Stefanie, and Elisa Izaurralde. 2015. "Towards a Molecular Understanding of MicroRNA-Mediated Gene Silencing." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3965.

Jones-Rhoades, Matthew W., David P. Bartel, and Bonnie Bartel. 2006. "MicroRNAs and Their Regulatory Roles in Plants." *Annual Review of Plant Biology*. https://doi.org/10.1146/annurev.arplant.57.032905.105218.

Kamenska, Anastasiia, Clare Simpson, Caroline Vindry, Helen Broomhead, Marianne Bénard, Michèle Ernoult-Lange, Benjamin P. Lee, Lorna W. Harries, Dominique Weil, and Nancy Standart. 2016. "The DDX6-4E-T Interaction Mediates Translational Repression and P-Body Assembly." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw565.

Kawamata, Tomoko, and Yukihide Tomari. 2010. "Making RISC." *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2010.03.009.

Khorshid, Mohsen, Jean Hausser, Mihaela Zavolan, and Erik Van Nimwegen. 2013. "A Biophysical MiRNA-MRNA Interaction Model Infers Canonical and Noncanonical Targets." *Nature Methods*. https://doi.org/10.1038/nmeth.2341.

Khvorova, Anastasia, Angela Reynolds, and Sumedha D. Jayasena. 2003. "Functional SiRNAs and MiRNAs Exhibit Strand Bias." *Cell* 115 (2): 209–16. https://doi.org/10.1016/S0092-8674(03)00801-8.

Kim, Doyeon, You Me Sung, Jinman Park, Sukjun Kim, Jongkyu Kim, Junhee Park, Haeok Ha, Jung Yoon Bae, Sohui Kim, and Daehyun Baek. 2016. "General Rules for Functional MicroRNA Targeting." *Nature Genetics*. https://doi.org/10.1038/ng.3694.

Kingston, Elena R., and David P. Bartel. 2019. "Global Analyses of the Dynamics of

Mammalian MicroRNA Metabolism." *BioRxiv*. https://doi.org/10.1101/607150.

Kiriakidou, M., Peter T. Nelson, Andrei Kouranov, Petko Fitziev, Costas Bouyioukos, Zissimos Mourelatos, and Artemis Hatzigeorgiou. 2004. "A Combined Computational-Experimental Approach Predicts Human MicroRNA Targets." *Genes & Development* 18 (10): 1165–78. https://doi.org/10.1101/gad.1184704.

Kleaveland, Benjamin, Charlie Y. Shi, Joanna Stefano, and David P. Bartel. 2018. "A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain." *Cell*. https://doi.org/10.1016/j.cell.2018.05.022.

Klum, Shannon M, Stanley D Chandradoss, Nicole T Schirle, Chirlmin Joo, and Ian J MacRae. 2018. "Helix-7 in Argonaute2 Shapes the MicroRNA Seed Region for Rapid Target Recognition." *The EMBO Journal*. https://doi.org/10.15252/embj.201796474.

Kozomara, Ana, Maria Birgaoanu, and Sam Griffiths-Jones. 2019. "MiRBase: From MicroRNA Sequences to Function." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1141.

Krek, Azra, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, et al. 2005. "Combinatorial MicroRNA Target Predictions." *Nature Genetics* 37 (5): 495–500. https://doi.org/10.1038/ng1536.

Kwak, Hojoong, Nicholas J. Fuda, Leighton J. Core, and John T. Lis. 2013. "Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing." *Science*. https://doi.org/10.1126/science.1229386.

La Mata, Manuel De, and Helge Großhans. 2018. "Turning the Table on MiRNAs." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/s41594-018-0040-x.

Lambert, Nicole, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. 2014. "RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins." *Molecular Cell* 54 (5): 887–900. https://doi.org/10.1016/j.molcel.2014.04.016.

Latreille, Mathieu, Jean Hausser, Ina Stützer, Quan Zhang, Benoit Hastoy, Sofia Gargani, Julie Kerr-Conte, et al. 2014. "MicroRNA-7a Regulates Pancreatic β Cell Function." *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI73066.

Lee, Yoontae, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, et al. 2003. "The Nuclear RNase III Drosha Initiates MicroRNA Processing." *Nature*. https://doi.org/10.1038/nature01957.

Lewis, Benjamin P., Christopher B. Burge, and David P. Bartel. 2005. "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets." *Cell* 120 (1): 15–20. https://doi.org/10.1016/j.cell.2004.12.035.

Lewis, Benjamin P., I. Hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. 2003. "Prediction of Mammalian MicroRNA Targets." *Cell* 115 (7): 787–98. https://doi.org/10.1016/S0092-8674(03)01018-3.

Lim, Lee P., Margaret E. Glasner, Soraya Yekta, Christopher B. Burge, and David P. Bartel. 2003. "Vertebrate MicroRNA Genes." *Science*. https://doi.org/10.1126/science.1080372.

Liu, Jidong, Michelle A. Carmell, Fabiola V. Rivas, Carolyn G. Marsden, J. Michael Thomson, Ji Joon Song, Scott M. Hammond, Leemor Joshua-Tor, and Gregory J. Hannon. 2004. "Argonaute2 Is the Catalytic Engine of Mammalian RNAi." *Science*. https://doi.org/10.1126/science.1102513.

Lovat, Francesca, Matteo Fassan, Pierluigi Gasparini, Lara Rizzotto, Luciano Cascione, Marco

Pizzi, Caterina Vicentini, et al. 2015. "MiR-15b/16-2 Deletion Promotes B-Cell Malignancies." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1514954112.

Lu, Li Fan, Mark P. Boldin, Ashutosh Chaudhry, Ling Li Lin, Konstantin D. Taganov, Toshikatsu Hanada, Akihiko Yoshimura, David Baltimore, and Alexander Y. Rudensky. 2010. "Function of MiR-146a in Controlling Treg Cell-Mediated Regulation of Th1 Responses." *Cell*. https://doi.org/10.1016/j.cell.2010.08.012.

Lu, Li Fan, Georg Gasteiger, I. Shing Yu, Ashutosh Chaudhry, Jing Ping Hsin, Yuheng Lu, Paula D. Bos, et al. 2015. "A Single Mirna-Mrna Interaction Affects The Immune Response In A Context- And Cell-Type-Specific Manner." *Immunity*. https://doi.org/10.1016/j.immuni.2015.04.022.

Martin, M. W., D. V. Grazhdankin, S. A. Bowring, D. A.D. Evans, M. A. Fedonkin, and J. L. Kirschvink. 2000. "Age of Neoproterozoic Bilatarian Body and Trace Fossils, White Sea, Russia: Implications for Metazoan Evolution." *Science*. https://doi.org/10.1126/science.288.5467.841.

Miska, Eric A., Ezequiel Alvarez-Saavedra, Allison L. Abbott, Nelson C. Lau, Andrew B. Hellman, Shannon M. McGonagle, David P. Bartel, Victor R. Ambros, and H. Robert Horvitz. 2007. "Most Caenorhabditis Elegans MicroRNAs Are Individually Not Essential for Development or Viability." *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.0030215.

Nguyen, Tuan Anh, Myung Hyun Jo, Yeon Gil Choi, Joha Park, S. Chul Kwon, Sungchul Hohng, V. Narry Kim, and Jae Sung Woo. 2015. "Functional Anatomy of the Human Microprocessor." *Cell*. https://doi.org/10.1016/j.cell.2015.05.010.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. "Expansion of the Eukaryotic Proteome by Alternative Splicing." *Nature*. https://doi.org/10.1038/nature08909.

Park, Chong Yon, Lukas T. Jeker, Karen Carver-Moore, Alyssia Oh, Huey Jiin Liu, Rachel Cameron, Hunter Richards, et al. 2012. "A Resource for the Conditional Ablation of MicroRNAs in the Mouse." *Cell Reports*. https://doi.org/10.1016/j.celrep.2012.02.008.

Park, Yongjin, and Manolis Kellis. 2015. "Deep Learning for Regulatory Genomics." *Nature Biotechnology*. https://doi.org/10.1038/nbt.3313.

Pfaff, Janina, Janosch Hennig, Franz Herzog, Ruedi Aebersold, Michael Sattler, Dierk Niessing, and Gunter Meister. 2013. "Structural Features of Argonaute-GW182 Protein Interactions." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1308510110.

Rehwinkel, Jan, Isabelle Behm-Ansmant, David Gatfield, and Elisa Izaurralde. 2005. "A Crucial Role for GW182 and the DCP1:DCP2 Decapping Complex in MiRNA-Mediated Gene Silencing." *RNA*. https://doi.org/10.1261/rna.2191905.

Reinhart, Brenda J., Frank J. Slack, Michael Basson, Amy E. Pasquienelll, Jill C. Bettlnger, Ann E. Rougvle, H. Robert Horvitz, and Gary Ruvkun. 2000. "The 21-Nucleotide Let-7 RNA Regulates Developmental Timing in Caenorhabditis Elegans." *Nature*. https://doi.org/10.1038/35002607.

Salomon, William E, Samson M Jolly, Melissa J Moore, Phillip D Zamore, William E Salomon, Samson M Jolly, Melissa J Moore, Phillip D Zamore, and Victor Serebrov. 2015. "Single-Molecule Imaging Reveals That Argonaute Reshapes the Binding Properties of Its Nucleic Acid Article Single-Molecule Imaging Reveals That Argonaute Reshapes the Binding

Properties of Its Nucleic Acid Guides." *Cell*. https://doi.org/10.1016/j.cell.2015.06.029.

Sanuki, Rikako, Akishi Onishi, Chieko Koike, Rieko Muramatsu, Satoshi Watanabe, Yuki Muranishi, Shoichi Irie, et al. 2011. "MiR-124a Is Required for Hippocampal Axogenesis and Retinal Cone Survival through Lhx2 Suppression." *Nature Neuroscience*. https://doi.org/10.1038/nn.2897.

Schirle, Nicole T., Jessica Sheu-Gruttadauria, Stanley D. Chandradoss, Chirlmin Joo, and Ian J. MacRae. 2015. "Water-Mediated Recognition of T1-Adenosine Anchors Argonaute2 to MicroRNA Targets." *ELife*. https://doi.org/10.7554/eLife.07646.

Schirle, Nicole T., Jessica Sheu-Gruttadauria, and Ian J. MacRae. 2014. "Structural Basis for MicroRNA Targeting." *Science*. https://doi.org/10.1126/science.1258040.

Schmiedel, Jörn M., Sandy L. Klemm, Yannan Zheng, Apratim Sahay, Nils Blüthgen, Debora S. Marks, and Alexander Van Oudenaarden. 2015. "MicroRNA Control of Protein Expression Noise." *Science*. https://doi.org/10.1126/science.aaa1738.

Schwanhüusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. "Global Quantification of Mammalian Gene Expression Control." *Nature*. https://doi.org/10.1038/nature10098.

Schwarz, Dianne S., György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D. Zamore. 2003. "Asymmetry in the Assembly of the RNAi Enzyme Complex." *Cell*. https://doi.org/10.1016/S0092-8674(03)00759-1.

Sood, Pranidhi, Azra Krek, Mihaela Zavolan, Giuseppe Macino, and Nikolaus Rajewsky. 2006. "Cell-Type-Specific Signatures of MicroRNAs on Target MRNA Expression." *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0511045103.

Subtelny, Alexander O., Stephen W. Eichhorn, Grace R. Chen, Hazel Sive, and David P. Bartel. 2014. "Poly(A)-Tail Profiling Reveals an Embryonic Switch in Translational Control." *Nature*. https://doi.org/10.1038/nature13007.

Suzuki, Hiroshi I., Akihiro Katsura, Takahiko Yasuda, Toshihide Ueno, Hiroyuki Mano, Koichi Sugimoto, and Kohei Miyazono. 2015. "Small-RNA Asymmetry Is Directly Driven by Mammalian Argonautes." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.3050.

Tsai, Wei Chih, Sheng Da Hsu, Chu Sui Hsu, Tsung Ching Lai, Shu Jen Chen, Roger Shen, Yi Huang, et al. 2012. "MicroRNA-122 Plays a Critical Role in Liver Homeostasis and Hepatocarcinogenesis." *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI63455.

Wee, Liang Meng, C. Fabián Flores-Jasso, William E. Salomon, and Phillip D. Zamore. 2012. "Argonaute Divides Its RNA Guide into Domains with Distinct Functions and RNA-Binding Properties." *Cell*. https://doi.org/10.1016/j.cell.2012.10.036.

Wei, Yusheng, Siwu Peng, Meng Wu, Ravi Sachidanandam, Zhidong Tu, Shihong Zhang, Christine Falce, Eric A. Sobie, Djamel Lebeche, and Yong Zhao. 2014. "Multifaceted Roles of MiR-1s in Repressing the Fetal Gene Program in the Heart." *Cell Research*. https://doi.org/10.1038/cr.2014.12.

Wu, Ligang, Jihua Fan, and Joel G. Belasco. 2006. "MicroRNAs Direct Rapid Deadenylation of MRNA." *Proceedings of the National Academy of Sciences of the United States of America* 103 (11): 4034–39. https://doi.org/10.1073/pnas.0510928103.

Xia, Tianbing, John SantaLucia, Mark E. Burkard, Ryszard Kierzek, Susan J. Schroeder, Xiaoqi

Jiao, Christopher Cox, and Douglas H. Turner. 1998. “Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson - Crick Base Pairs.” *Biochemistry*. https://doi.org/10.1021/bi9809425.

Yang, Edward, Erik van Nimwegen, Mihaela Zavolan, Nikolaus Rajewsky, Mark Schroeder, Marcelo Magnasco, and James E. Darnell. 2003. “Decay Rates of Human MRNAs: Correlation with Functional Characteristics and Sequence Attributes.” *Genome Research*. https://doi.org/10.1101/gr.1272403.

Zhang, Haidi, Fabrice A. Kolb, Lukasz Jaskiewicz, Eric Westhof, and Witold Filipowicz. 2004. “Single Processing Center Models for Human Dicer and Bacterial RNase III.” *Cell*. https://doi.org/10.1016/j.cell.2004.06.017.

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 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.

Abstract

MicroRNAs (miRNAs) act within Argonaute proteins to guide repression of mRNA 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 ≤12-nucleotide sequences. This approach revealed noncanonical target sites unique 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-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 target (Jonas and Izaurralde 2015). miRNAs are grouped into families based on 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 (Wee et al. 2012; Salomon et al. 2015; Schirle, Sheu-Gruttadauria, and MacRae 2014; Schirle et al. 2015; Jo et al. 2015; Klum et al. 2018; Chandradoss et al. 2015). 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 inference through correlative approaches. These models focus on mRNAs with canonical 6–8-nt sites matching the miRNA seed region (Fig. 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), using datasets that monitor mRNA changes that occur after introducing a miRNA (Grimson et al. 2007; Agarwal et al. 2015; Gumienny and Zavolan 2015; Paraskevopoulou et al. 2013). Although the correlative model implemented in TargetScan7 performs as well as the best in vivo crosslinking 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 ($r^2 = 0.14$) (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.

Results

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 (Lambert et al. 2014; Dominguez et al. 2018). 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 (Fig. 1B), we purified human AGO2 loaded with miR-1 (Flores-Jasso, Salomon, and Zamore 2013) (fig. 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 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 RNA library (fig. 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 (perfect match to miR-1 nucleotides 2–8, enrichment 14 fold), then the 7mer-A1 site (perfect match to miR-1 nucleotides 2–7 followed by an A, enrichment 7.2 fold), and the 6mer site (perfect match to miR-1 nucleotides 2–7, enrichment 3.0 fold) (Fig. 1, A and C). Little if any enrichment was observed for either the 6mer-A1 site (perfect match to miR-1 nucleotides 2–6 followed by an A) or the 6mer-m8 site (perfect match to miR-1 nucleotides 3–8) at this lowest concentration of 7.3 pM AGO2–miR-1 (Fig. 1, A and C), consistent with their weak signal in previous analyses of conservation and efficacy(Friedman et al. 2009; Agarwal et al. 2015; 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 (fig. S1D). Although sites with supplementary pairing can have enhanced efficacy and affinity (Bartel, 2009; Brennecke, Stark, Russell, & Cohen, 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 pM to 73 pM, enrichment for each of the six site types increased (Fig. 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 (Fig. 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 (Fig. 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 ≥0.994 for each of the ten pairwise comparisons, fig. S1, F and G). These quantitative binding affinities followed the same hierarchy as observed for site enrichment, but the differences in affinities were of greater magnitude (Fig. 1D and fig. 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) for the ten most enriched *k*-mers, determined the extent of complementarity to 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 (Fig. 1, E and F).

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 (Fig. 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 4 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, as 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 (Fig. 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 (Fig. 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, Hannon, and Darnell 2012; Loeb et al. 2012; Helwak et al. 2013; Khorshid et al. 2013; Grosswendt et al. 2014), 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; Wee et al. 2012; Salomon et al. 2015; Chi, Hannon, and Darnell 2012; Loeb et al. 2012), and lsy-6 of nematodes, which is thought to bind unusually weakly to its canonical sites (Garcia et al. 2011) (Fig. 2 and fig. S2, B and C). In the case of let-7a, previous biochemical analyses have determined the K_D values of a few sites (Wee et al. 2012; Salomon et al. 2015), and our values agreed well, which further validated our high-throughput approach (fig. S1H).

The site-affinity profile of let-7a resembled that of miR-1, except the 6mer-m8 and 6mer-A1 site for let-7a had greater binding affinity than essentially all of the noncanonical sites (Fig. 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 (Figs. 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, (Fig. 2, B and C, and fig. S2, B and C). 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 (fig. 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 (Fig. 2, B and C, and fig. 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 (Fig. 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 (Fig. 3A).

When evaluating other types of noncanonical sites proposed to confer widespread repression in mammalian cells (Kim et al. 2016; Chi, Hannon, and Darnell 2012), 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 (fig. S3). The other was the pivot site (Chi, Hannon, and Darnell 2012), which was bound by miR-124 (e.g., 8mer-bG(6.7); Fig. 2C) and lsy-6 (e.g., 8mer-bA(6.7); fig. S2B) but not by the other four miRNAs (fig. S4). Thus, 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 striking 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. Although prior studies show that AGO proteins remodel the thermodynamic properties of their loaded RNA guides (Wee et al. 2012; Salomon et al. 2015), 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 (Fig. 3B, left). The apparent binding-energy contributions of the m8 and A1 ($\Delta\Delta G_{\mathrm{m8}}$ and $\Delta\Delta G_{\mathrm{A1}}$, respectively) were largely independent, as inferred from the relative K_{D} values of the four site types. That is, for each miRNA, the $\Delta\Delta G_{\mathrm{m8}}$ 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 (Fig. 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 (Fig. 1A) (Schirle et al. 2015; Lewis, Burge, and Bartel 2005). 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 (Fig. 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 $K_D = e^{-\Delta G/RT}$ (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 sites to the transfected miRNA, we observed a striking relationship between AGO-RBNS–determined K_D values and mRNA fold-changes (Fig. 3, D to I, 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 (Fig. 3, F and G). 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 (Fig. 3, D to I). These included the pivot sites for miR-124 and lsy-6, the AA-6mer-m8 site for miR-124, and the bulge-G7-containing sites for lsy-6 and miR-7 (Fig. 3, G to I).

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 (fig. 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 (fig. 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, Burge, and Bartel 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, based on the dinucleotide sequences immediately flanking each side of the

8mer, and determined relative K_D values for each (Fig. 4A). This analysis revealed a ~100-fold range in values, depending on the identities of the flanking dinucleotides, with binding affinity strongly tracking the AU content of the flanking dinucleotides. Extending this analysis across all miR-1 site types (Fig. 4B), as well as to sites to the other five miRNAs (fig. S6, A to E), 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 (Fig. 4B and fig. S6, A to E).

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 (Fig. 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 (fig. S6F). A and U nucleotides each enhanced affinity, whereas G nucleotides reduced affinity, and C nucleotides were intermediate or neutral (Fig. 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, Sheu-Gruttadauria, and MacRae 2014), contributed ~2-fold more to binding affinity than did the 3′ flanking sequence (Fig. 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. 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 the difference was greatest for the samples with the lower AGO2–miR-1 concentrations (fig. 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 based on 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 (Fig. 4D and fig. S6H). A striking 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 (fig. S6I, $r^2 = 0.79$). Taken together, our results demonstrated that local sequence context has a large influence on miRNA–target binding affinity, and indicated that this influence results predominantly from the differential propensities of flanking sequences to form 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 (Fig. 3, D to I), 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 (Denzler et al. 2016; Bosson, Zamudio, and Sharp 2014; Jens and Rajewsky 2015) 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 (Fig. 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 (Figs. 1F and 2). For each mRNA *m* and transfected miRNA *t*, the steady-state occupancy $N_{m,t}$ (i.e., average number of AGO–miRNA complexes loaded with miRNA *t* bound to mRNA *m*) was predicted as a function of the K_D values of the potential binding sites contained within the mRNA ORF and 3′ UTR, as well as the concentration of the unbound AGO–miRNA$_t$ complex a_t, which was fit as a single value for each transfected miRNA (Fig. 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 $N_{m,t}$ (Fig. 5B, equation 2). To isolate the effects of a transfected miRNA over background, we further offset our prediction of repression by a background binding term (Fig. 5B, $N_{m,t,\text{ background}}$).

The calculation of predicted repression required an estimate of how much a single bound RISC complex affected the mRNA decay rate (Fig. 5B, *b*), which was fit as a global value. Additionally, to account for the observation that sites in open reading frames (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 (Fig. 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 let-7 has high endogenous expression in HeLa cells). A strong correspondence was observed when comparing mRNA changes measured upon miRNA transfection to those predicted by the model (fig. S7A, $r^2 = 0.30–0.37$).

The overall performance of our biochemical model ($r^2 = 0.34$, Fig. 5C) 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-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 (fig. 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 (Fig. 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/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 (Grimson et al. 2007; Agarwal et al. 2015) and allowed them to modify the K_D values linearly 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 (fig. S7C). Simultaneously incorporating all three metrics to generate what we call the "biochemical+ model" improved the r^2 by 9% to 0.37 (Fig. 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 out-performed TargetScan7 (Fig. 5G. This improvement extended to predicting repression after transfecting miR-124 and miR-7 into HEK293 cells (Hausser et al. 2009) (fig. 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 (fig. S8, B to D). Furthermore, for individual CLIP clusters enriched upon miR-155 knockout, we observed a strong relationship between the cluster occupancy predicted by our K_D values and the observed enrichment of the cluster ($r_s = 0.46$, $p < 10^{-7}$, fig. S8E), supporting the conclusion that K_D values measured in vitro reflect intracellular AGO binding.

When provided 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 (Fig. 5F, green bars, $r^2 = 0.35$), and conversely when provided K_D values for only the 12-nt *k*-mers without a canonical site, the model still retained some predictive power (Fig. 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 (Fig. 5F, light-blue bars) and when considering only the noncanonical sites (Fig. 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 (Fig. 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.

Convolutional neural network 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; Jaganathan et al. 2019; Tunney et al. 2018; Cuperus et al. 2017), we chose a convolutional neural network (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 (fig. S9A), and the second was the previously described biochemical model that links intracellular repression with relative K_D values (Fig. 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 × 12 × 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 (fig. S9B), the model considered both strands of each transfected duplex, which allowed the neural network to learn from another 16 AGO-loaded guide sequences.

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 (fig. 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 (Fig. 3, D to I), striking 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 (Fig. 6B and fig. 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 crosslinking results (Khorshid et al. 2013) (Fig. 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 uniquely able to stratify sites of the same type to different miRNAs (e.g., Fig. 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 (fig. 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 54% improvement over TargetScan, and performance dropped slightly when either shuffling or omitting noncanonical sites (Fig. 6D and fig. 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 (Fig. 6D). The CNN-predicted relative K_D values also enabled the biochemical+ model to outperform TargetScan and crosslinking approaches in predicting the effects of deleting or adding a miRNA in other cellular contexts (fig. S12, B to D).

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 (Fig. 5F and fig. 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 (fig. S12E, median of five datasets), and when provided with CNN-predicted values it explained ~50% of the variability attributable to direct targeting (fig. S12F, median of twelve datasets).

Insights into miRNA targeting

Our results provide new insight into both the canonical and noncanonical miRNA site types. For each miRNA, the canonical 8mer site was the highest-affinity site identified, illustrating its primacy in miRNA targeting. However, the canonical 7mer-m8 was the not always the second-most effective site; miR-155 had one noncanonical site with greater affinity than that of this canonical site, and miR-124 had three (Fig. 2, B and C). Moreover, four of the six miRNAs had noncanonical sites with greater affinity than that of the canonical 7mer-A1 sites. Indeed, miR-124 had 25 noncanonical sites with greater affinity than that of the canonical 7mer-A1 site and 33 noncanonical sites with greater affinity than that of the canonical 6mer site. (Fig. 2 and fig. S2).

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, all six canonical site types were identified for five of the six miRNAs (the sole 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 (Figs. 1F and 2 and fig. S2, B and C). 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. For example, although the miR-124 7mer-m8 site had lower affinity than a 11-nt noncanonical site, the canonical 7-nt site occupied much more AGO2–miR-124 because of its

256-fold greater abundance. 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, Burge, and Bartel 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 such pairing can reduce site affinity (Salomon et al. 2015), another reports that it slightly 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 (Fig. 4), whereas pairing to nucleotides 9 and 10 enhanced affinity in the context of 3′-only sites (Fig. 2, C and D). 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). Moreover, we found that although the nucleotides at target positions 9 and 10 seem unable to pair to the miRNA in the context of most canonical sites, nucleotide composition at positions 9 and 10 can have a dramatic influence on the affinity of canonical sites through an effect on site accessibility (Fig. 4).

The success of our biochemical model in predicting how each mRNA would respond to a transfected miRNA (Fig. 5) supports the conclusion that site binding affinity is the major determinant of miRNA-mediated repression and that noncanonical sites measurably contribute to this repression in the cell. 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_t, i.e., 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_t was fit separately for each miRNA and passenger strand of each transfection experiment. Nonetheless, when a_t values were deviated from the fitted values, the biochemical+ model still outperformed TargetScan in predicting test-set repression over a 100-fold range of values (Fig. 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_t values should come with the determination of these values for more miRNAs in more cellular contexts, together with the observation that, as expected, fitted a_t values are higher for miRNAs with lower predicted target abundance and lower general affinity for their targets (Fig. 6F).

Our work replaced the correlative models of targeting efficacy with a principled biochemical model that explains and predicts about half 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, i.e., the widespread differences in site preferences observed for different miRNAs and the striking 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 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.

Figures and figure legends

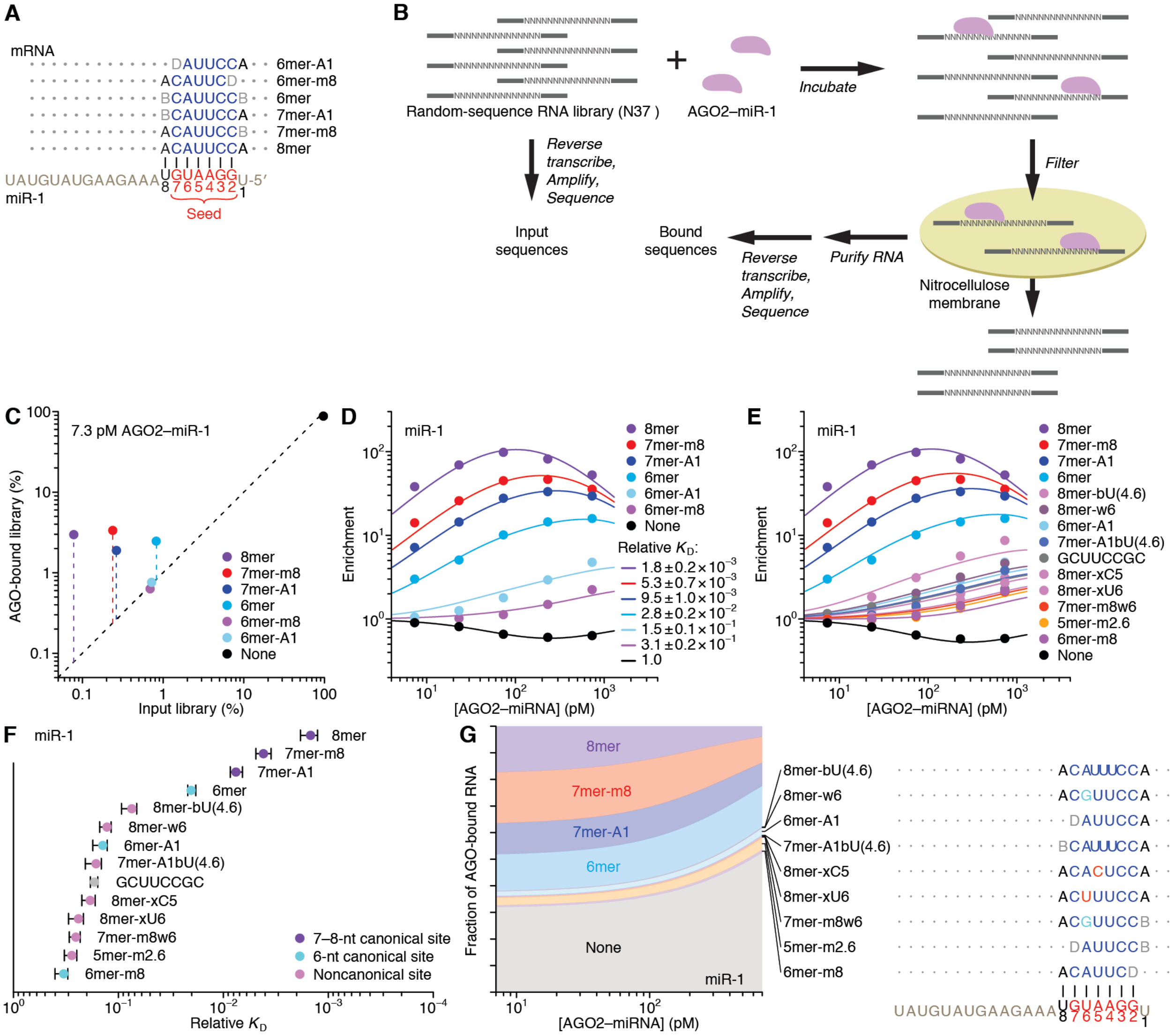


Fig. 1. AGO-RBNS reveals binding affinities of canonical and novel miR-1 target sites. (**A**) Canonical sites of miR-1. These sites have contiguous pairing (blue) to the miRNA seed (red), and some include an additional match to miRNA nucleotide 8 or an A opposite miRNA nucleotide 1 (B, not A; D, not C). (**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 to 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. Otherwise, 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, 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 colors of (E). At 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 non-complementarity (gray; B, not A; D, not C; H, not G; V, not U). The GCUUCCGC motif is omitted because it did not match miR-1 and did not mediate repression by miR-1 (fig. S5B).

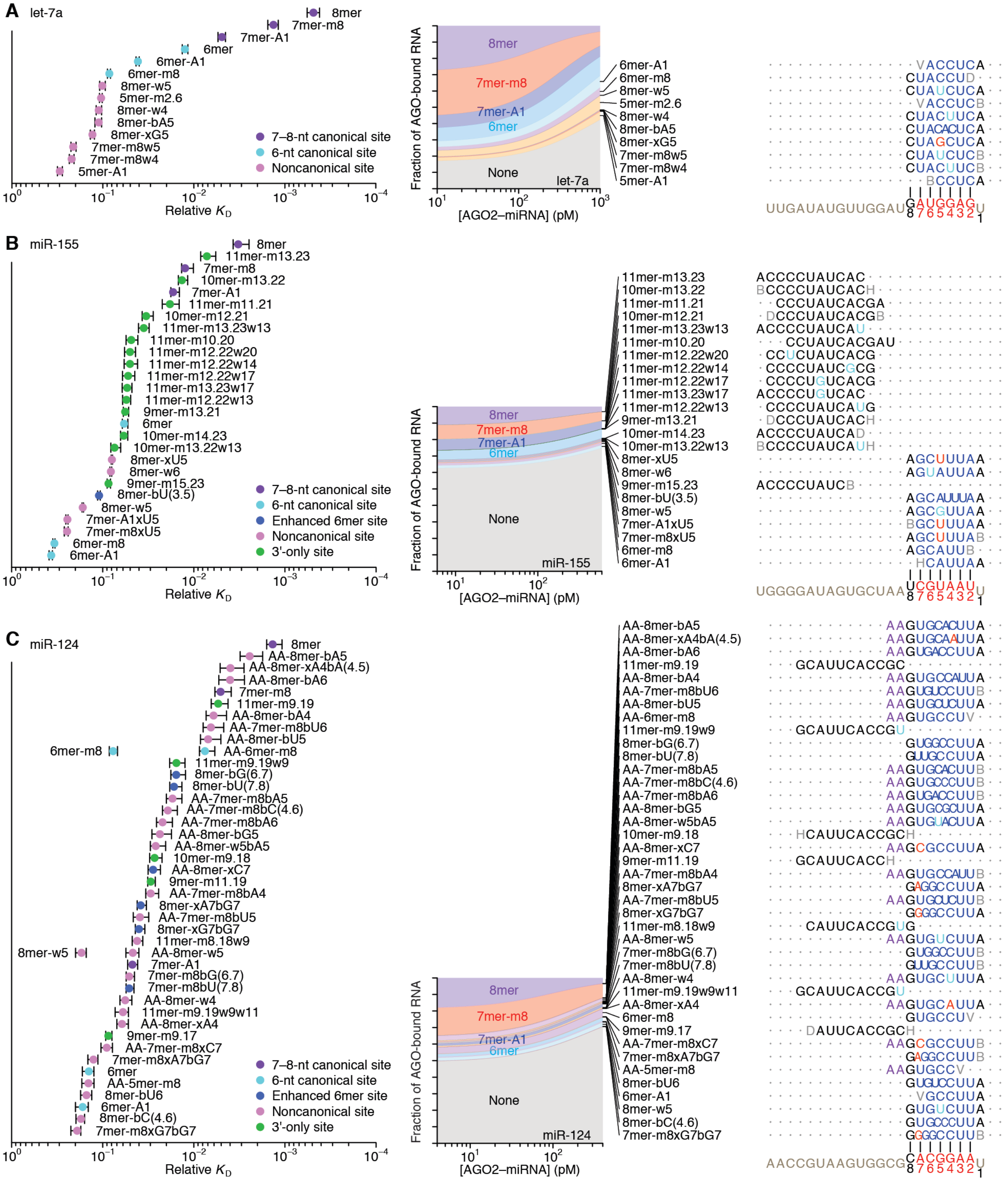


Fig. 2. Distinct canonical and noncanonical binding of different miRNAs. (**A** to **C**) Relative K_D values and proportional occupancy of established and newly identified sites of let-7a (A), miR-155 (B), and miR-124 (C). The two miR-124 sites that were present as a 5′-AA-extended

form in addition to an unextended form are shown on the same line (C). Relative K_D values are plotted as in Fig. 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 bugled nucleotide (blue) (B and C). The proportion of AGO2–miRNA bound to each site type is estimated and shown as in Fig. 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 (fig. S5B).

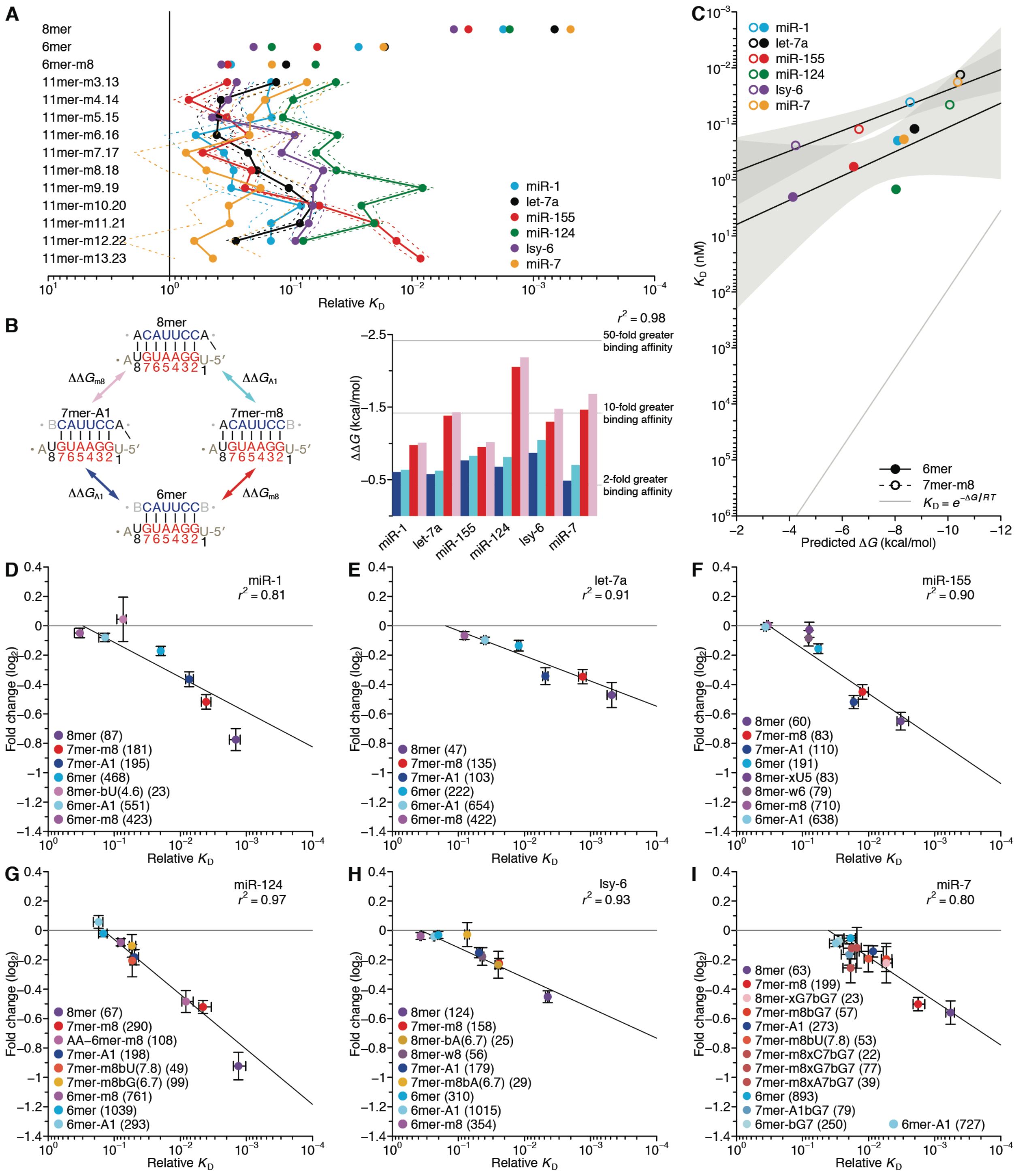


Fig. 3. Additional analyses of binding affinities and the correspondence between binding affinity and repression efficacy. (A) Diverse functionality and position dependence of 11-nt 3′-only sites. Relative K_D values for each potential 11-nt 3′-only site are plotted for the indicated

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 Fig. 1F. The solid and dashed lines indicate geometric mean and 95% confidence interval, respectively, determined as in Fig. 1D. (**B**) The independent contributions of the A1 and m8 features. At 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. At 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** to **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 Figs. 1 and 2 and fig. 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.

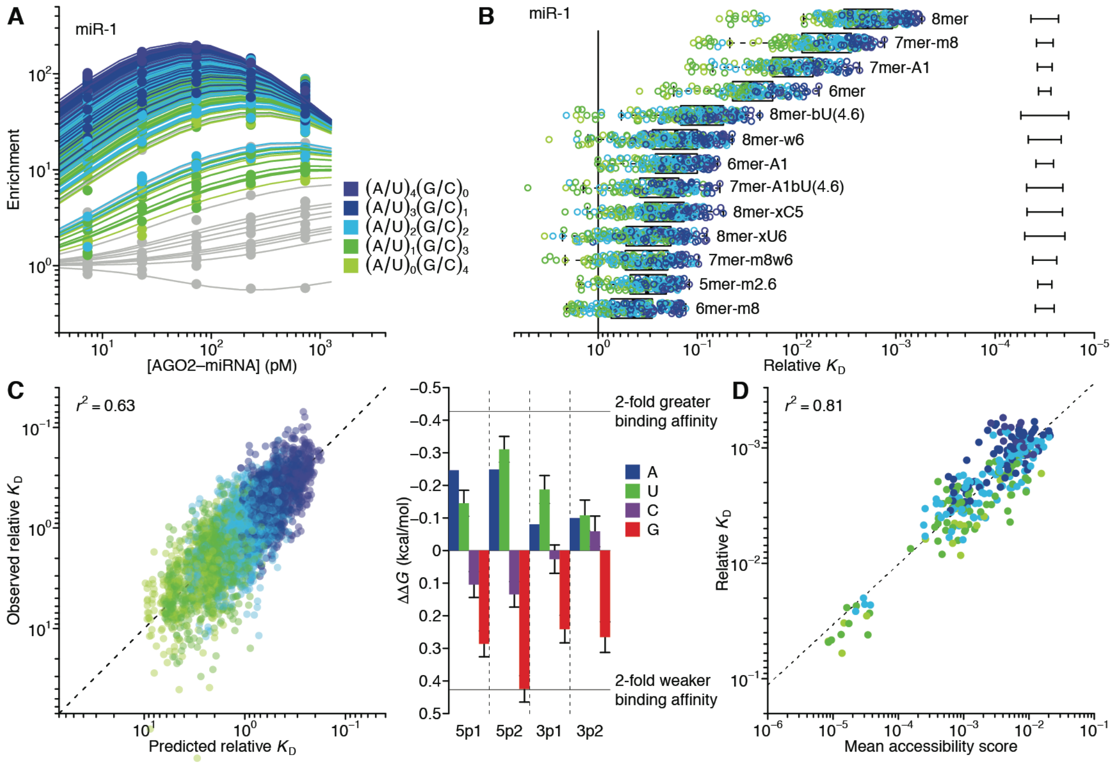


Fig. 4. The influence of flanking dinucleotide sequence context. (**A**) AGO-RBNS profile of miR-1 sites, showing results for the 8mer separated into 256 different 12-nt sites based on the identities of the two dinucleotides immediately flanking the 8mer. For each 12-nt site, the points and line are colored based on the AU content of the flanking dinucleotides (key). For context, results of Fig. 1E are re-plotted in gray. Otherwise as in Fig. 1E. (**B**) Relative K_D values for the each miR-1 site identified in Fig. 1F separated into 144–256 sites as in (A) based on the identities of the flanking dinucleotides. The points are colored as in (A). Error bars, median 95% confidence interval across all K_D values. Otherwise, as in Fig. 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 six-fold cross validation, training on the results for five miRNAs and reporting the predictions for the held-out miRNA. The r^2 quantifies the agreement between the predicted and actual values. At 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 fig. S6G.

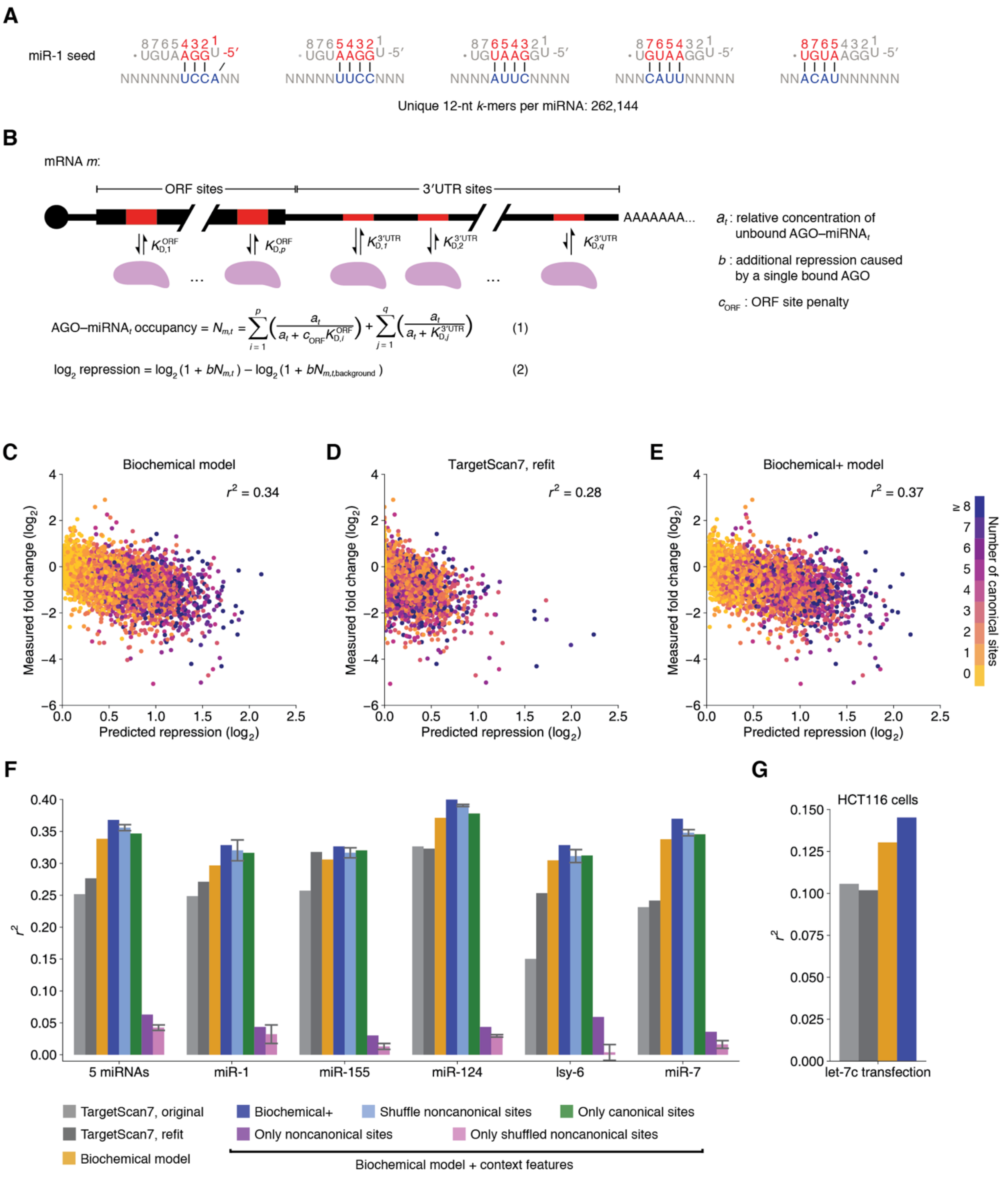


Fig. 5. AGO-RBNS K_D values enable a predictive model of miRNA-mediated repression in cells. (**A**) The 262,144 12-nt *k*-mers with at least four contiguous matches to the extended seed region of miR-1, for which relative K_D values were determined. Relative K_D values were

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. Otherwise, as in (C). (**E**) Performance of the biochemical+ model. Otherwise, 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_t fit for the five miRNAs in (F). Otherwise, as in (F).

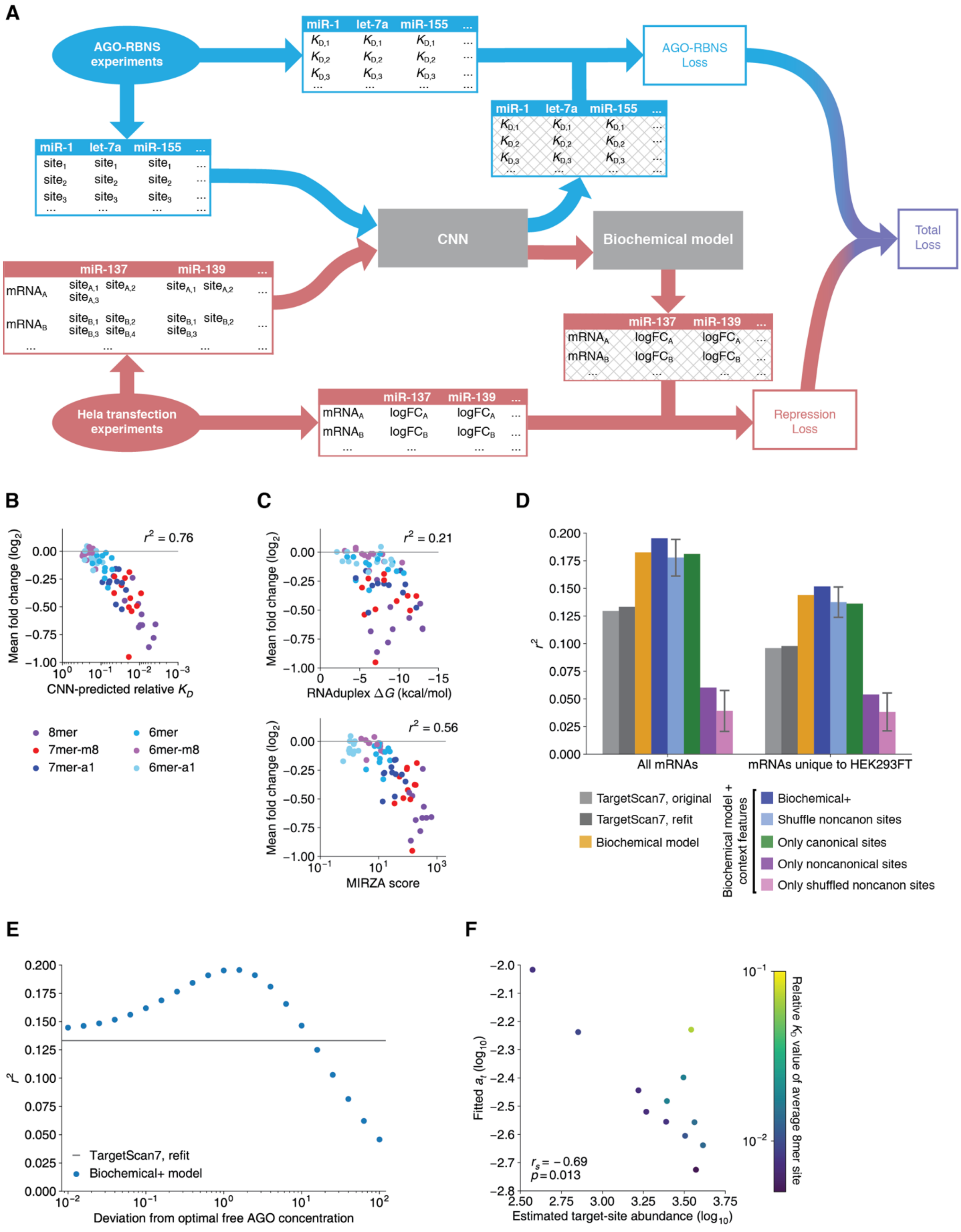

A
AGO-RBNS experiments
miR-1 let-7a miR-155 ...
$K_{D,1}$ $K_{D,2}$ $K_{D,3}$
AGO-RBNS Loss
site$_1$ site$_2$ site$_3$
CNN
Biochemical model
Total Loss
miR-137 miR-139 ...
mRNA$_A$ mRNA$_B$
site$_{A,1}$ site$_{A,2}$ site$_{A,3}$
site$_{B,1}$ site$_{B,2}$ site$_{B,3}$ site$_{B,4}$
logFC$_A$ logFC$_B$
Hela transfection experiments
Repression Loss
B
r^2 = 0.76
Mean fold change (log$_2$)
0.00 −0.25 −0.50 −0.75 −1.00
10^0 10^{-1} 10^{-2} 10^{-3}
CNN-predicted relative K_D
8mer
7mer-m8
7mer-a1
6mer
6mer-m8
6mer-a1
C
r^2 = 0.21
0 −5 −10 −15
RNAduplex ΔG (kcal/mol)
r^2 = 0.56
10^{-1} 10^1 10^3
MIRZA score
D
r^2
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
All mRNAs
mRNAs unique to HEK293FT
TargetScan7, original
TargetScan7, refit
Biochemical model
Biochemical model + context features
Biochemical+
Shuffle noncanon sites
Only canonical sites
Only noncanonical sites
Only shuffled noncanon sites
E
r^2
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
10^{-2} 10^{-1} 10^0 10^1 10^2
Deviation from optimal free AGO concentration
TargetScan7, refit
Biochemical+ model
F
Fitted a_t (log$_{10}$)
−2.0 −2.1 −2.2 −2.3 −2.4 −2.5 −2.6 −2.7 −2.8
2.50 2.75 3.00 3.25 3.50 3.75
Estimated target-site abundance (log$_{10}$)
r_s = −0.69
p = 0.013
10^{-1} 10^{-2}
Relative K_D value of average 8mer site

Fig. 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. Tables with hash marks indicate model-predicted values, rather than experimentally measured values. (**B**) The relationship between repression efficacy and CNN-predicted relative K_D values for the canonical sites for the 12 test miRNAs. Otherwise, as in Fig. 3D to I. (**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. Otherwise, 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. Otherwise, this panel is as in Fig. 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 deviating the a_t values away from the optimal fitted values. (**F**) Relationship between fitted a_t and estimated target-site abundance (Garcia et al. 2011) for the guide strands of the 12 transfected miRNA duplexes. Points are colored by the average relative K_D value of the 8mer site to each miRNA. The Spearman *r* and p value for the relationship are shown.

Methods

Experimental methods

1 Purification of AGO2–miRNA complexes

5′ phosphorylated RNAs of each miRNA duplex 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, Salomon, and Zamore 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 di-alanine spacer. 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, Lebovitz, and Roeder 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 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, 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 μ L of the sample spotted onto a Hybond nylon (Thermo Fisher, 45001147) filter membrane alongside 1 μ L 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 (fig. 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 (section 10). 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 (section 10).

2 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, were added for inter-library comparison, and radiolabeled 18- and 30-nt standards were added for size selection. After gel purification on a 15% polyacrylamide urea gel, RNA was ligated to a pre-adenylated 3′ adapter 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 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 the Illumina HiSeq platform with 40-nt single reads. To count the miRNAs in each library, the sequence corresponding to the first 18 nucleotides following the 5′ adapter were queried against a list of the first 18 nucleotides of human miRNAs annotated in miRbase_v21, 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 fig. S1A.

3 Preparation of RNA libraries for AGO-RBNS

Four libraries of DNA oligonucleotides, each containing a central region of 37 random-sequence positions, 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, 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-^{32}P]-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.

4 Preparation of AGO-RBNS quantification standards

Defined RNAs were added to each AGO-RBNS sequencing library at the step of the Proteinase K incubation (section 5) to enable quantitative comparison of the RNA recovered in each binding sample. These quantification standards were generated by in vitro transcription of the corresponding PCR templates followed by TURBO DNase treatment, gel purification, CIP treatment, and gel purification, as described for the RNA libraries (section 3).

5 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 (section 3), 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 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.

6 miRNA transfections and mRNA-seq library preparation

RNAs of each miRNA duplex 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.

7 Massively parallel reporter library design and preparation

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.

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. 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) 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. 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).

8 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 (section 5) 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, 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.

Computational and mathematical methods

9 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 small-RNA 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 a TCG at positions 38–40 in the case of the first AGO2–miR-1 experiment, or a TGT at these positions for all other experiments.

10 De novo site identification

To identify sites of an AGO–miRNA complex using RBNS results, we performed an analysis in which we repeatedly 1) calculated the enrichment of all 10-nt *k*-mers in the sample library corresponding to 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, in which each experiment used a separately purified AGO–miRNA complex (section 1).

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 that 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. 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 (section 1), 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 top10-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 (Fig. 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 (Fig. 4 and fig. 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 (Fig. 2C).

11 Determination of K_D values from AGO-RBNS data

11.1 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

$$\ln \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \ln p(\boldsymbol{y} \mid \boldsymbol{x}(\boldsymbol{\theta})), \tag{11.1.1}$$

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

$$f_{grad}(\boldsymbol{\theta}) = \nabla f_{cost}(\boldsymbol{x}(\boldsymbol{\theta})). \tag{11.1.2}$$

The optimization routine was performed with the *optim* function in R (R Foundation for Statistical Computing. 2018) 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.

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

The function $\boldsymbol{x}(\boldsymbol{\theta})$ produces an *m*×*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 *m*th site type) and the total number of binding reactions comprising that AGO-

RBNS experiment (which was 5 for all experiments), respectively. 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 ***l*** is estimated using

$$\boldsymbol{l} = \frac{\boldsymbol{y}^l}{\sum_{i=1}^{m} y_i^l} \times 100 \text{ nM}, \tag{11.2.1}$$

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

$$\begin{aligned} \boldsymbol{a} &= a \times \boldsymbol{s} \\ &= a \times (0.4\%,\ 1.27\%,\ 4\%,\ 12.7\%,\ 40\%), \end{aligned} \tag{11.2.2}$$

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

$$(K_1,\ K_2,\ \ldots,\ K_m,\ a,\ b). \tag{11.2.3}$$

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 $\boldsymbol{\theta}$:

$$\begin{aligned} K_1 &= e^{\theta_1} \\ &\vdots \\ K_m &= e^{\theta_m} \\ a &= e^{\theta_{m+1}} \\ b &= e^{\theta_{m+2}}, \end{aligned} \tag{11.2.4}$$

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

$$x_{ij} = c_{ij} + g_{ij}, \tag{11.2.5}$$

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

$$g_{ij} = \alpha_j l_{ij}^f, \tag{11.2.6}$$

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

$$\begin{aligned} \sum_{i=1}^{m} g_{ij} &= b \\ \sum_{i=1}^{m} \alpha_j l_{ij}^f &= b \\ \alpha_j &= \frac{b}{\sum_{i=1}^{m} l_{ij}^f}. \end{aligned} \tag{11.2.7}$$

Substituting for α_j in equation (11.2.6) using equation (11.2.7), and further substituting for g_{ij} in equation (11.2.5) yields

$$x_{ij} = c_{ij} + \frac{b}{\sum_{i'=1}^{m} l_{i'j}^{f}} l_{ij}^{f}. \qquad (11.2.8)$$

By invoking the conservation of mass for each site type (i.e., $c_{ij} + l_{ij}^{f} = l_i$), equation (11.2.8) can be expressed as

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

$$x_{ij} = c_{ij} \left(1 - \frac{b}{L - C_j} \right) + l_i \frac{b}{L - C_j}, \qquad (11.2.9)$$

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 (11.2.9) 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:

$$K_i \equiv \frac{a_j^f l_{ij}^f}{c_{ij}}, \qquad (11.2.10)$$

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

$$K_i = \frac{a_j^f (l_i - c_{ij})}{c_{ij}}, \tag{11.2.11}$$

which is rearranged to give

$$c_{ij} = \frac{l_i a_j^f}{a_j^f + K_i}. \tag{11.2.12}$$

Using equation (11.2.12) to substitute for c_{ij} in equation (11.2.9) yields

$$x_{ij} = l_i \left(\frac{a_j^f}{a_j^f + K_i} \left(1 - \frac{b}{L - C_j} \right) + \frac{b}{L - C_j} \right), \tag{11.2.13}$$

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

$$x_{ij} = l_i \left(\frac{a_j^f}{a_j^f + K_i} \left(1 - \frac{b}{L - \sum_{i'=1}^{m} \frac{l_i a_j^f}{a_j^f + K_{i'}}} \right) + \frac{b}{L - \sum_{i'=1}^{m} \frac{l_i a_j^f}{a_j^f + K_{i'}}} \right). \tag{11.2.14}$$

This is the final form of the function, wherein read abundances are modeled from the fixed vector ***l*** (and its sum *L*) and the parameter vector $\boldsymbol{\theta}$ where $K_i = e^{\theta_i}$ for $i \in [1 \mathinner{..} m]$, $a = e^{\theta_{m+1}}$, and and $b = e^{\theta_{m+2}}$, and whose values are iteratively updated during the optimization routine. Equation (11.2.14) 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*:

$$a_j = a_j^f + \sum_{i=1}^{m} c_{ij}. \tag{11.2.15}$$

Because each c_{ij} value is itself a function of ***l***, ***K***, and *a* according to equation (11.2.12), equation (11.2.15) 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

$$f(a_j^f) = as_j - a_j^f - \sum_{i=1}^{m} \frac{l_i a_j^f}{a_j^f + K_i} \tag{11.2.16}$$

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.

11.3 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*

$$\begin{aligned} f_{\text{cost}}(\boldsymbol{x}) &= -\ln \prod_{j=1}^{n} f_{\text{mult}}(\boldsymbol{y}_j, \boldsymbol{\pi}_j) \\ &= -\ln \prod_{j=1}^{n} \frac{Y_j! \prod_{i=1}^{m} \pi_{ij}^{y_{ij}}}{\prod_{i=1}^{m} y_{ij}!}, \end{aligned} \tag{11.3.1}$$

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

$$f_{\text{cost}}(\boldsymbol{x}) = \sum_{j=1}^{n} \left(Y_j \ln X_j - \sum_{i=1}^{m} y_{ij} \ln x_{ij} + \sum_{i=1}^{m} \ln y_{ij}! - \ln Y_j! \right). \tag{11.3.2}$$

After discarding the third and fourth terms in equation (11.3.2) 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

$$f_{\text{cost}}(\boldsymbol{x}) = \sum_{j=1}^{n} \left(Y_j \ln X_j - \sum_{i=1}^{m} y_{ij} \ln x_{ij} \right). \tag{11.3.3}$$

11.4 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 $\boldsymbol{\theta}$:

$$\begin{aligned} f_{\text{grad}}(\boldsymbol{\theta}) &= \nabla f_{\text{cost}}(\boldsymbol{x}(\boldsymbol{\theta})) \\ &= \left(\frac{\partial f_{\text{cost}}}{\partial \theta_1}, \frac{\partial f_{\text{cost}}}{\partial \theta_2}, \ldots, \frac{\partial f_{\text{cost}}}{\partial \theta_m} \right). \end{aligned} \tag{11.4.1}$$

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:

$$\begin{aligned}\frac{df_{\text{cost}}}{d\theta_k} &= \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{\partial f_{\text{cost}}}{\partial x_{ij}}\frac{dx_{ij}}{d\theta_k} \\ &= \sum_{j=1}^{n}\sum_{i=1}^{m}\frac{\partial f_{\text{cost}}}{\partial x_{ij}}\left(\frac{\partial x_{ij}}{\partial\theta_k}+\sum_{i'=1}^{m}\frac{\partial x_{ij}}{\partial c_{i'j}}\frac{dc_{i'j}}{d\theta_k}\right).\end{aligned} \tag{11.4.2}$$

$\frac{\partial f_{\text{cost}}}{\partial x_{ij}}$ is obtained by differentiating equation (11.3.3)

$$\frac{\partial f_{\text{cost}}}{\partial x_{ij}} = \frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}, \tag{11.4.3}$$

and both $\frac{\partial x_{ij}}{\partial\theta_k}$ and $\frac{\partial x_{ij}}{\partial c_{i'j}}$ are obtained by differentiation of equation (11.2.9)

$$\begin{aligned}\frac{\partial x_{ij}}{\partial\theta_k} &= e^{\theta_k}\frac{l_i - c_{ij}}{L - C_j}\delta_{k(m+2)} \\ &= b\frac{l - c_{ij}}{L - C_j}\delta_{k(m+2)},\end{aligned} \tag{11.4.4}$$

$$\frac{\partial x_{ij}}{\partial c_{i'j}} = b\frac{l_i - c_{ij}}{(L - C_j)^2} + \left(1 - \frac{b}{L - C_j}\right)\delta_{i'i}, \tag{11.4.5}$$

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

$$\delta_{ab} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \tag{11.4.6}$$

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 (11.4.2) using (11.4.3), (11.4.4), and (11.4.5), respectively, and rearranging yields

$$\frac{df_{\text{cost}}}{d\theta_k} = \sum_{j=1}^{n} \frac{1}{L - C_j} \sum_{i=1}^{m} \left(\left(\frac{Y_i}{X_j} - \frac{y_{ij}}{x_{ij}} \right) \times \right.$$

$$\left. \left(b(l_i - c_{ij})\delta_{k(n+2)} + (L - C_j - b)\frac{dc_{ij}}{d\theta_k} + b\frac{l_i - c_{ij}}{L - C_j}\frac{dC_j}{d\theta_k} \right) \right). \qquad (11.4.7)$$

Inspection of equation (11.4.7) 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 (11.4.7) 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

$$\frac{dc_{ij}}{d\theta_k} = \frac{\partial c_{ij}}{\partial \theta_k} + \frac{\partial c_{ij}}{\partial a_j^f}\frac{da_j^f}{d\theta_k}, \qquad (11.4.8)$$

and differentiation of equation (11.2.15) yields

$$as_j\delta_{k(m+1)} = \frac{da_j^f}{d\theta_k} + \sum_{i=1}^{m}\frac{dc_{ij}}{d\theta_k}. \qquad (11.4.9)$$

Substituting for $\frac{da_j^f}{d\theta_k}$ in equation (11.4.8) with equation (11.4.9) results in

$$\frac{dc_{ij}}{d\theta_k} = \frac{\partial c_{ij}}{\partial \theta_k} + \frac{\partial c_{ij}}{\partial a_j^f}\left(as_j\delta_{k(m+1)} - \sum_{i=1}^{m}\frac{dc_{ij}}{d\theta_k} \right), \qquad (11.4.10)$$

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 (11.4.10) for all site types $i \in [1 \,..\, m]$ yields

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

$$\frac{dC_j}{d\theta_k} = \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial \theta_k} + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f} as_j \delta_{k(m+1)} - \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f} \frac{dC_j}{d\theta_k}, \tag{11.4.11}$$

Rearranging equation (11.4.11) yields

$$\frac{dC_j}{d\theta_k} = \frac{\sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial \theta_k} + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f} as_j \delta_{k(m+1)}}{1 + \sum_{i=1}^{m} \frac{\partial c_{ij}}{\partial a_j^f}} \tag{11.4.12}$$

For the purposes of clarity, we define

$$\phi_{ij} \equiv \frac{\partial c_{ij}}{\partial a_j^f} = \frac{l_i K_i}{(a_j^f + K_i)^2}, \tag{11.4.13}$$

such that

$$\frac{\partial c_{ij}}{\partial \theta_k} = \frac{-a_j^f l_i K_i}{(a_j^f + K_i)^2} \delta_{ki} = -a_j^f \phi_{ij} \delta_{ki}, \tag{11.4.14}$$

and we also define

$$\Phi_j \equiv \sum_{i=1}^{m} \phi_{ij}. \tag{11.4.15}$$

Equation (11.4.12) now reads as

$$\frac{dC_j}{d\theta_k} = \frac{-a_j^f \phi_{kj} \mathbb{I}_{[1..m]}(k) + \Phi_j as_j \delta_{k(m+1)}}{1+\Phi_j}, \tag{11.4.16}$$

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

$$\mathbb{I}_{[a..b]}(x) = \begin{cases} 1 & \text{if } x \in [a \,..\, b], \\ 0 & \text{if } x \notin [a \,..\, b]. \end{cases} \tag{11.4.17}$$

Substituting for $\frac{dC_j}{d\theta_k}$ into equation (11.4.10) using equation (11.4.16) yields

$$\begin{aligned} \frac{dc_{ij}}{d\theta_k} &= -a_j^f \phi_{ij} \delta_{ki} + \phi_{ij} \left(as_j \delta_{k(m+1)} - \frac{-a_j^f \phi_{kj} \mathbb{I}_{[1..m]}(k) + \Phi_j as_j \delta_{k(m+1)}}{1+\Phi_j} \right) \\ &= -a_j^f \phi_{ij} \left(\delta_{ki} - \frac{\phi_{kj} \mathbb{I}_{[1..m]}(k)}{1+\Phi_j} \right) + \frac{\phi_{ij} as_j \delta_{k(m+1)}}{1+\Phi_j}. \end{aligned} \tag{11.4.18}$$

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 (11.4.7) using equations (11.4.18) and (11.4.16), 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 (section 9.5).

11.5 Parameter initialization for relative K_D estimation

Each θ_i where $i \in [1, .., \mathrm{m}]$ (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:

$$\theta_{i,0} = \ln\left(\frac{1}{n}\sum_{j=1}^{n}\frac{y_{ij}}{Y_j}\Bigg/\frac{y_i^l}{Y^l}\right), \qquad (11.5.1)$$

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 (Figs 1 to 4 , and figs. S1 to S6) and 12-nt *k*-mers (Figs. 5 and 6, and figs. S7 to S12), respectively.

11.6 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 $j = \mathrm{ceil}\left(\frac{i}{40}\right)$ withheld from matrix $\boldsymbol{y}$, and with the read counts in the input sequencing $\boldsymbol{y}^I$ 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 is withheld during bootstrapping. To overcome this, we first performed the optimization using all five 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 (section 11.4) to 0.

11.7 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. 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 5 nucleotides of constant primer-binding sequence. 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**AAGTGTCCTT**, which has an 8mer site [GTGCCTT**A**] overlapping an AA-7mer-m8bU6 site [**AAGTGTCCTT**]) were not included, such that the procedure for estimating K_D values used only reads containing single sites. When analyzing the relative affinity of all possible 11-nt registers of pairing (Fig. 3A), of sites identified in Kim *et al*. (Kim et al. 2016) (fig. S3), or of sites with all possible single-nucleotide bulges and deletions (fig. 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 (Figs. 5 and 6, and figs. S7 to 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).

11.8 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 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.

12 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

$$K_{ijk} = \exp\left(s_{ij} + \sum_{p=1}^{4} \beta_p(n_{kp}) + \sum_{p=1}^{2} \gamma_p(d_{kp}) \right), \tag{12.1.1}$$

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 Fig. 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 Fig. 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).

13 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 wth root. The parameter w (and therefore the value after the –u flag) was either set to 15 in previous studies

(Fig. 4D and figs. S6, G and I) (Agarwal et al. 2015) or was allowed to span a range of values from 0 to 30 (fig. S6H).

14 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. Transcript annotations were from (Agarwal et al. 2015). 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 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:

$$\log \mathrm{TPM}_{m,t,b} = \beta_{m,b} + \beta_{m,t}. \tag{14.1.1}$$

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

15 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 (14.1.1) as per those of the HeLa transfection experiments.

16 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 (Fig. 3, D to I, and fig. S6F). This was done by first calculating the repression of each mRNA *m* by miRNA *t* as

$$r_{m,t} = \beta_{m,t} - \overline{\beta_{m^*,t}}, \tag{16.1.1}$$

where $\beta_{m,t}$ is its batch-normalized expression of in units of logTPM (section 14), 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

$$r_{m,t} = \sum_{j=1}^{N} n_{m,t,j} c_j, \tag{16.1.2}$$

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.

17 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 (Figs. 5 and 6) were calculated as described (section 9) 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.

18 Biochemical model for predicting repression

18.1 Modeling AGO occupancy and mRNA repression

Given the free concentration of miRNA-loaded AGO2, a_t, 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

$$\theta_{m,\mathrm{UTR3}} = \frac{a_t}{a_t + K_D}. \tag{18.1.1}$$

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:

$$\theta_{m,\mathrm{ORF}} = \frac{a_t}{a_t + c_{\mathrm{ORF}} K_D}. \tag{18.1.2}$$

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

$$N_{m,t} = \sum_{i=1}^{p} \frac{a_t}{a_t + c_{\mathrm{ORF}} K_{D,i}} + \sum_{j=1}^{q} \frac{a_t}{a_t + K_{D,j}}. \tag{18.1.3}$$

For a given mRNA *m* and miRNA *t* in a transfection experiment, let $N_{m,t}$ 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, $y_{m,t}$, according to its transcription rate and aggregate decay rate:

$$\frac{dy_{m,t}}{dt} = \alpha_m - \beta_m (1 + bN_{m,t}) y_{m,t}. \tag{18.1.4}$$

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

$$y_{m,t} = \frac{\alpha_m}{\beta_m (1 + bN_{m,t})}. \tag{18.1.5}$$

If the mRNA were not bound by the transfected miRNA at all (i.e., $N_{m,t} = 0$), its steady-state abundance would be

$$y_{m,0} = \frac{\alpha_m}{\beta_m}. \tag{18.1.6}$$

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

$$r_{m,t} = \frac{y_{m,t}}{y_{m,0}} = \frac{1}{1+bN_{m,t}}. \tag{18.1.7}$$

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:

$$\log r_{m,t} = -\log(1+bN_{m,t}). \tag{18.1.8}$$

18.2 Fitting the biochemical model to RNA-seq measurements

We could not measure $y_{m,0}$, and thus $r_{m,t}$ directly. However, 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 *t* out of *T* miRNA transfection experiments,

$$\begin{aligned}\log r_{m,t} - \overline{\log \mathbf{r}_m} &= (\log y_{m,t} - \log y_{m,0}) - \frac{1}{T}\sum_{i=1}^{T}(\log y_{m,i} - \log y_{m,0}) \\ &= (\log y_{m,t} - \log y_{m,0}) - \frac{1}{T}\sum_{i=1}^{T}\log y_{m,i} + \frac{1}{T}\sum_{i=1}^{T}\log y_{m,0} \\ &= (\log y_{m,t} - \log y_{m,0}) - \frac{1}{T}\sum_{i=1}^{T}\log y_{m,i} + \log y_{m,0} \\ &= \log y_{m,t} - \frac{1}{T}\sum_{i=1}^{T}\log y_{m,i} \\ &= \log y_{m,t} - \overline{\log \boldsymbol{y}_m}. \end{aligned} \tag{18.2.1}$$

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

$$L = \sum_{m=1}^{M}\sum_{t=1}^{T}((\log y_{m,t} - \overline{\log \mathbf{y}_m}) - \log(\log \hat{r}_{m,t} - \overline{\log \hat{\mathbf{r}}_m}))^2. \tag{18.2.2}$$

These values were used to calculate the r^2 values. For plotting, 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 (Fig. 5, C to E) 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.

18.3 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.

18.4 Calculating site occupancy in the biochemical+ model

All the additional features modified in 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,

$$\log K_{D,i,\mathrm{biochem+}} = \log K_{D,i} + c_{\mathrm{SA}}\mathrm{SA}_i + c_{\mathrm{Threep}}\mathrm{Threep}_i + c_{\mathrm{PCT}}\mathrm{PCT}_i, \tag{18.4.1}$$

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

18.5 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 (18.2.2) using the 16 transfection experiments of miRNA duplexes into HeLa cells.

19 Combined CNN and biochemical model

19.1 CNN architecture

The CNN architecture was as described in fig. 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.

19.2 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 10 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:

$$L_{\text{rbns}} = \sum_{i=1}^{10} (\log K_{D,i} - \log \hat{K}_{D,i})^2. \tag{19.2.1}$$

For each of the 32 miRNAs, 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 12-nt matches to the same miRNA 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 (18.2.2). Here, *t* enumerates the 16 miRNAs in the training set, *m* enumerates the 16 mRNAs in the mini-batch, and $n_{m,t}^{\text{guide,ORF}}$, $n_{m,t}^{\text{pass,ORF}}$, $n_{m,t}^{\text{guide,3}'\text{UTR}}$, and $n_{m,t}^{pass\text{,3}'\text{UTR}}$

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 *t*:

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

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

$$r_{m,t} = \frac{1}{1 + b(\theta_{m,t}^{\text{guide,ORF}} + \theta_{m,t}^{\text{guide,UTR3}} + \theta_{m,t}^{\text{pass,ORF}} + \theta_{m,t}^{\text{pass,ORF}})} \tag{19.2.2}$$

$$L_{\text{repression}} = \sum_{m=1}^{16} \sum_{t=1}^{16} ((\log y_{m,t} - \overline{\log \mathbf{y}_m}) - (\log \hat{r}_{m,t} - \overline{\log \hat{\mathbf{r}}_m}))^2. \tag{19.2.3}$$

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_{t} values, an L_2 regularization term was also applied to the differences between guide-strand a_{t} values and the average guide-strand a_{t} value to prevent these values from drifting too far apart initially.

$$\mathbf{d}^{\text{guide}} = \mathbf{a}_t^{\text{guide}} - \overline{\mathbf{a}^{\text{guide}}} \tag{19.2.4}$$

The RBNS loss weight, repression loss weight, CNN weight regularizer, and the a_{t} offset weights are λ_k, λ_r, λ_w, and λ_d respectively.

$$L_{\text{total}} = \lambda_k L_{\text{rbns}} + \lambda_r L_{\text{repression}} + \lambda_w (\|\mathbf{w}_1\|_2 + \|\mathbf{w}_2\|_2 + \|\mathbf{w}_3\|_2 + \|\mathbf{w}_4\|_2) + \lambda_d \|\mathbf{d}^{\text{guide}}\|_2). \tag{19.2.5}$$

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.

19.3 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 sequences, and sorted each group. If the 32 sequences in a group 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 section 16.

19.4 Predictions of miRNA–target interaction energy using other methods

To calculate the free-energy of binding for canonical site-types to each miRNA (Fig. 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.

20 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 in (Agarwal et al. 2015). For gene expression changes upon knockout of miR-122 in mouse liver cells, 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). 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 (section 17) were used, whereas CNN-predicted *K*D values were used for the miR-302/367, miR-122, and miR-20a datasets using mRNA sequences from miRbase release 22 (Kozomara, Birgaoanu, and Griffiths-Jones 2019). When predicting mRNA changes upon miR-155 knockout in mouse T cells, the average a_t value of passenger strands fit for the HeLa transfection datasets was used. For all other datasets, the average a_t value of miRNA strands fit for the HeLa transfection datasets was used.

21 Estimation of maximal r^2 values

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

X: Actual 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 $\boldsymbol{w} = [w_1, w_2 ... w_m]$ with probabilities $\boldsymbol{p} = [p_1, p_2 ... p_m]$. In practice, we used 50 bins spanning –3 to 0 in log space (–4.33 to 0 in log2 space). To calculate the probability of observing the measured repression values $\boldsymbol{y}_{1,2...n} \sim Y$ given $(\sigma_1^2 + \sigma_2^2)$, ***w***, and ***p***

$$\begin{aligned}\log p(\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{w}, \sigma_1, \sigma_2) &= \log \prod_{i=1}^{n} p(y_i \mid \boldsymbol{p}, \boldsymbol{w}, \sigma_1, \sigma_2) \\ &= \sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{p}, \boldsymbol{w}, \sigma_1, \sigma_2) \\ &= \sum_{i=1}^{n} \log(\sum_{j=1}^{m} p_j \tfrac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-(y_i - w_j)^2 / 2(\sigma_1^2+\sigma_2^2)}).\end{aligned} \tag{21.1.1}$$

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

$$\sigma_2^2 \approx Var(Y_{rep1} - Y_{rep2}) / 2 \tag{21.1.2}$$

$$\sigma_1^2 = (\sigma_1^2 + \sigma_2^2) - \sigma_2^2 \tag{21.1.3}$$

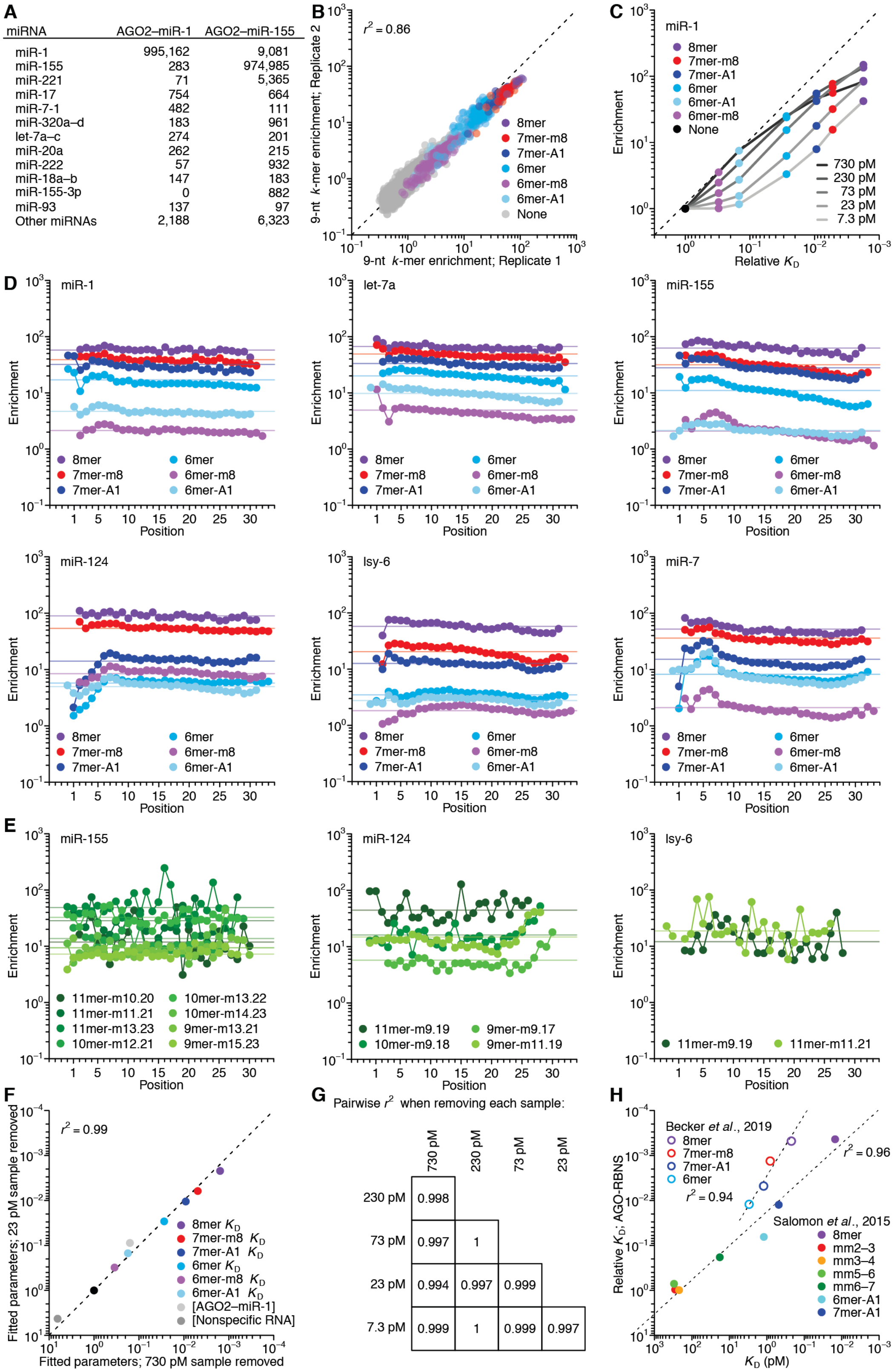
and estimated the expected value and variance of X given $\boldsymbol{w}$, and $\boldsymbol{p}$:

$$E(X) \approx \frac{1}{m}\sum_{j=1}^{m} p_j w_j \tag{21.1.4}$$

$$Var(X) \approx \frac{1}{m}\sum_{j=1}^{m} p_j (w_j - E(X))^2 \tag{21.1.5}$$

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

Supplementary figures and legends



miRNA	AGO2–miR-1	AGO2–miR-155
miR-1	995,162	9,081
miR-155	283	974,985
miR-221	71	5,365
miR-17	754	664
miR-7-1	482	111
miR-320a–d	183	961
let-7a–c	274	201
miR-20a	262	215
miR-222	57	932
miR-18a–b	147	183
miR-155-3p	0	882
miR-93	137	97
Other miRNAs	2,188	6,323

Pairwise r^2 when removing each sample:

	730 pM	230 pM	73 pM	23 pM
230 pM	0.998			
73 pM	0.997	1		
23 pM	0.994	0.997	0.999	
7.3 pM	0.999	1	0.999	0.997

Fig. 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 (Fig. 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 (Salomon et al. 2015; Becker et al. 2019). 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. 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) 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 (Fig. 2A).

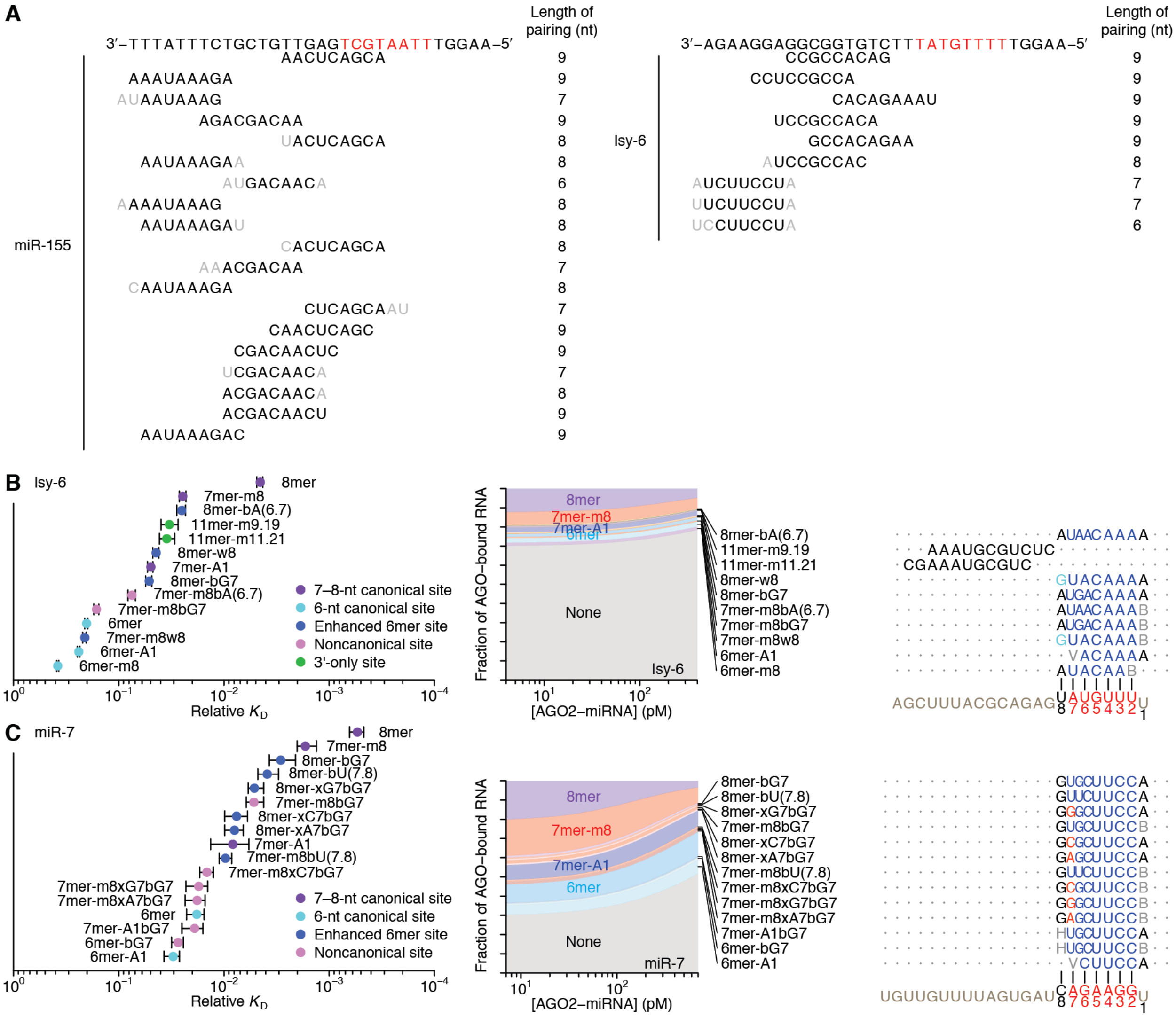


Fig. S2. Additional sites identified through AGO-RBNS. (**A**) Enriched motifs that were identified for miR-155 and lsy-6 yet lacked complementarity to the respective guide sequence, aligned to highlight their complementarity to the competitor oligo used to purify the AGO–miRNA complex. Because these 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 Fig. 2. 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 (fig. S5B).

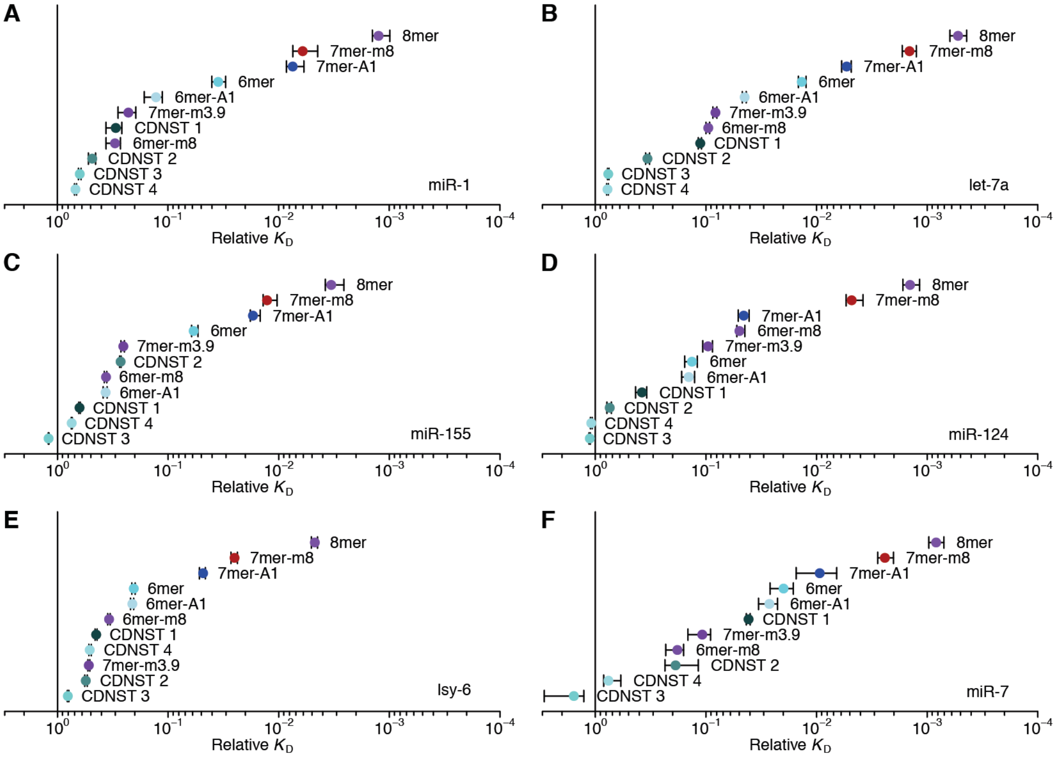


Fig. S3. Relative K_D values of site types reported in Kim *et al.* (2016). (**A** to **E**) Analysis was as in Fig. 1F but performed using the site types of (Kim et al. 2016), which include the canonical sites (Fig. 1A), an offset 7mer (which pairs to miRNA nucleotides 3–9), as well as four context-dependent 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 (Fig. 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 (Figs. 1F and 2A). This site was also detected in analysis of our first miR-7 replicate. CDNST 2 is a 7mer-A1 site with a mismatch at position 5; this site includes the 7mer-A1xU5 site identified for miR-155 (Fig. 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.

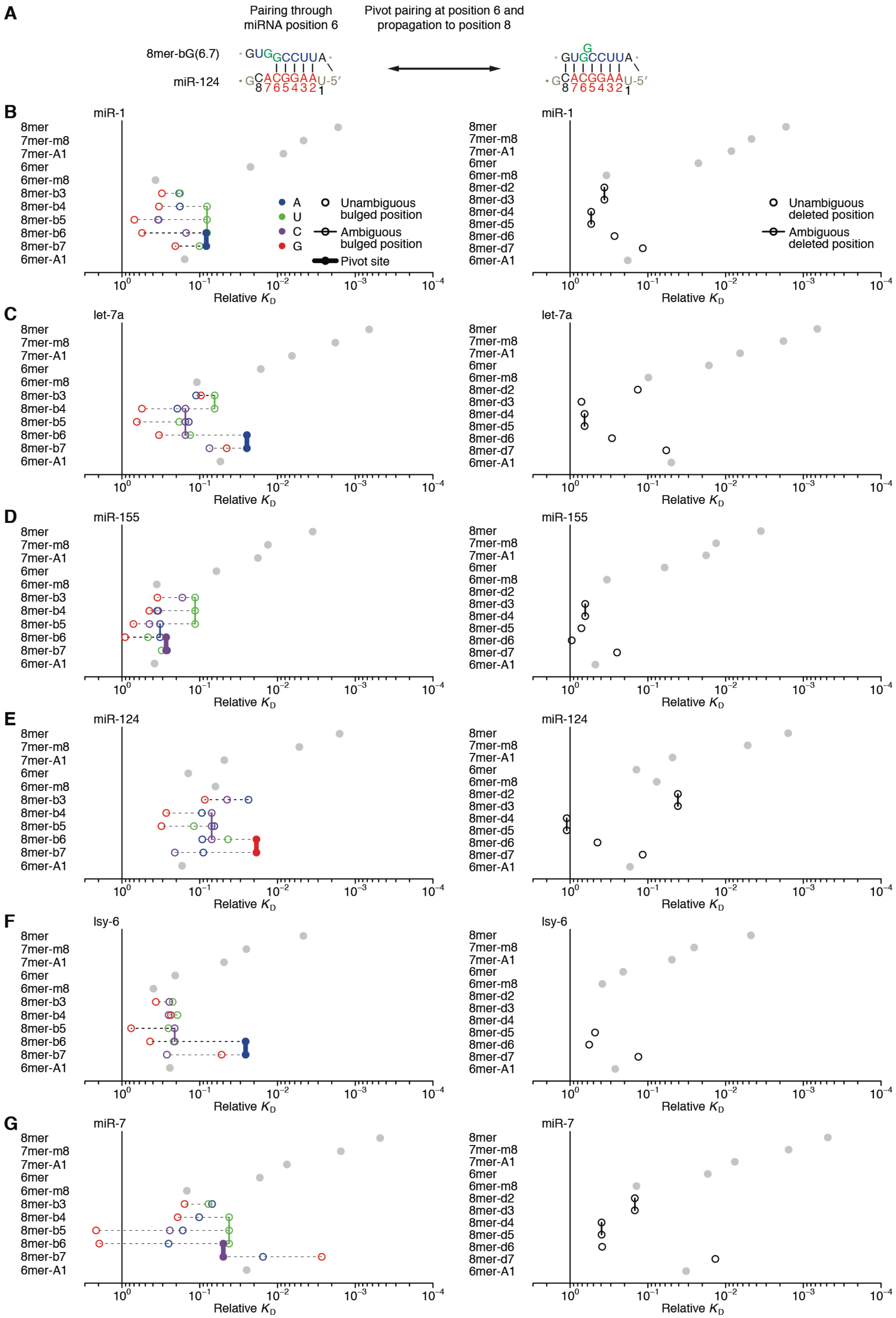

A
Pairing through miRNA position 6
Pivot pairing at position 6 and propagation to position 8
8mer-bG(6.7)
GUGGCCUUA
miR-124
GCACGGAAU-5′
87654321
G
GUGCCUUA
GCACGGAAU-5′
87654321
B
miR-1
8mer
7mer-m8
7mer-A1
6mer
6mer-m8
8mer-b3
8mer-b4
8mer-b5
8mer-b6
8mer-b7
6mer-A1
A
U
C
G
Unambiguous bulged position
Ambiguous bulged position
Pivot site
8mer-d2
8mer-d3
8mer-d4
8mer-d5
8mer-d6
8mer-d7
Unambiguous deleted position
Ambiguous deleted position
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Relative K_D
C
let-7a
D
miR-155
E
miR-124
F
lsy-6
G
miR-7

Fig. 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, Hannon, and Darnell 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** to **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 Fig. 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, Hannon, and Darnell 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 (Fig. 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 (Fig. 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, Hannon, and Darnell 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, Hannon, and Darnell 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 (Fig. 2C and fig. 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 (Figs. 1F, and 2, A and B, and fig. 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 (fig. S4). Bulged nucleotides within the miRNA strand abrogated binding, presumably due to steric constraints imposed by AGO.

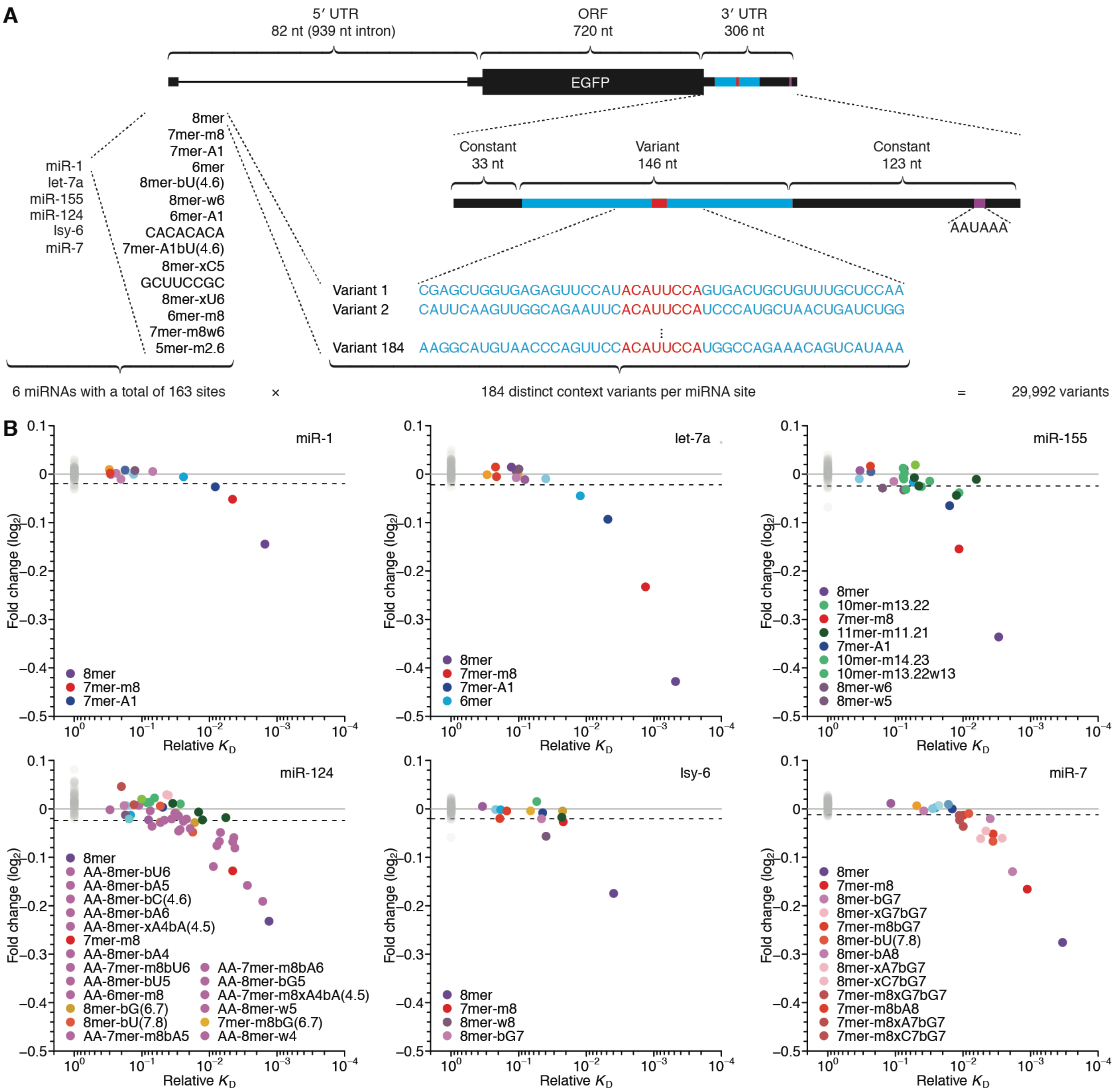


Fig. S5. Massively parallel reporter assay to monitor the effects of sites identified by AGO-RBNS. (**A**) Schematic of the EGFP pre-mRNA expressed upon transfection of the library of reporter plasmids. The top, middle, and bottom diagrams respectively depict the pre-mRNA, the 3′ UTR, and a region within 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).

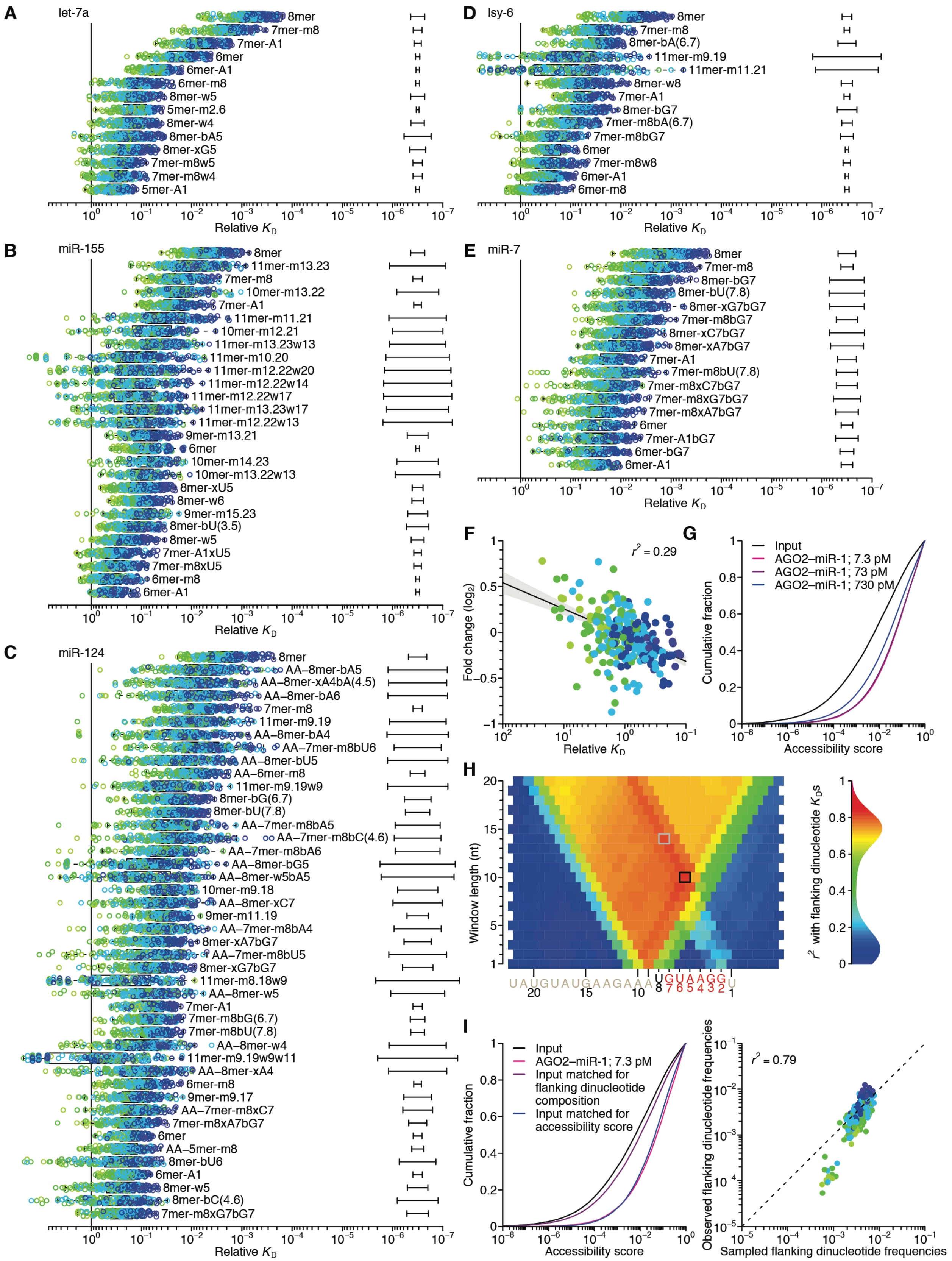

A
let-7a
8mer
7mer-m8
7mer-A1
6mer
6mer-A1
6mer-m8
8mer-w5
5mer-m2.6
8mer-w4
8mer-bA5
8mer-xG5
7mer-m8w5
7mer-m8w4
5mer-A1
Relative K_D
B
miR-155
8mer
11mer-m13.23
7mer-m8
10mer-m13.22
7mer-A1
11mer-m11.21
10mer-m12.21
11mer-m13.23w13
11mer-m10.20
11mer-m12.22w20
11mer-m12.22w14
11mer-m12.22w17
11mer-m13.23w17
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
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
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 ($\log_2$)
1
0.5
0
–0.5
–1
10^2
10^1
10^0
10^{-1}
Relative K_D
G
Input
AGO2–miR-1; 7.3 pM
AGO2–miR-1; 73 pM
AGO2–miR-1; 730 pM
Cumulative fraction
0
0.2
0.4
0.6
0.8
1
10^{-8}
10^{-6}
10^{-4}
10^{-2}
10^0
Accessibility score
H
Window length (nt)
20
15
10
5
1
UAUGUAUGAAGAAAUGUAAGGU
20
15
10
8
7 6 5 4 3 2
1
r^2 with flanking dinucleotide K_Ds
1
0.8
0.6
0.4
0.2
0
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
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}

Fig. S6. The influence of flanking dinucleotide context. (**A** to **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 Fig. 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 144–256 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 Fig. 4C, and they *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 Fig. 3, D to I (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 (Fig. 4D and fig. 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 (fig. 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 (fig. S6I, right).

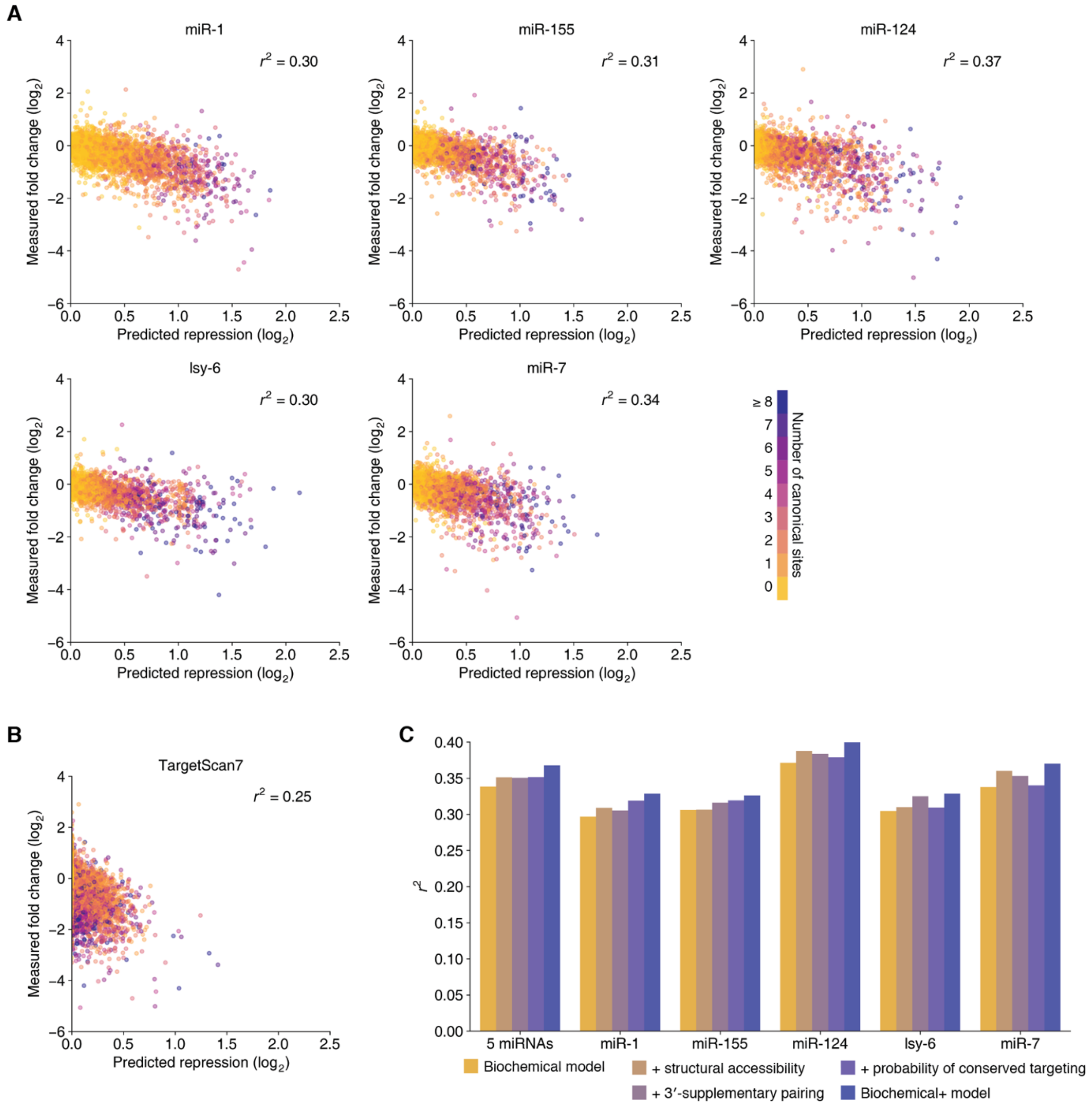


Fig. S7. **Additional analyses of the biochemical models.** (**A**) Performance of the biochemical model as evaluated for each of the five miRNAs individually. Otherwise, as in Fig. 5C. (**B**) Performance of the published version of the TargetScan7 model as evaluated using the combined results of five 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.

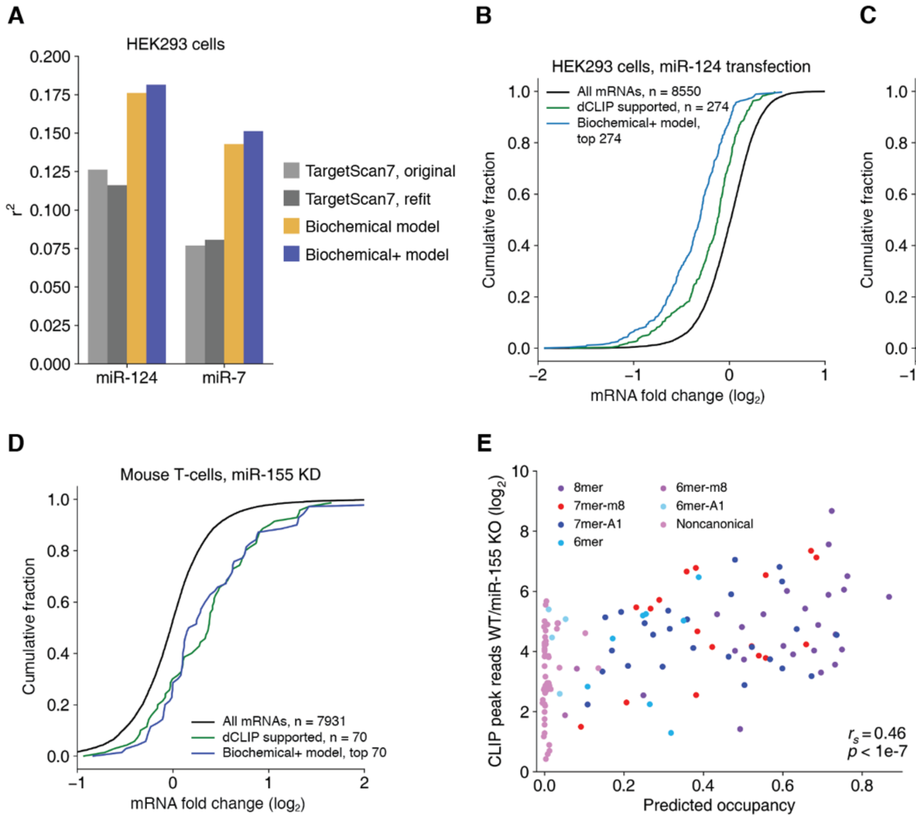


Fig. S8. **Evaluation of the biochemical models using other published datasets.** (**A**) Performances of the biochemical and biochemical+ models compared to those of both the published and refit versions of TargetScan7, as evaluated using mRNA fold changes observed after transfecting either miR-124 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 differential photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) (Hafner et al. 2010) (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 miR-155–AGO 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).

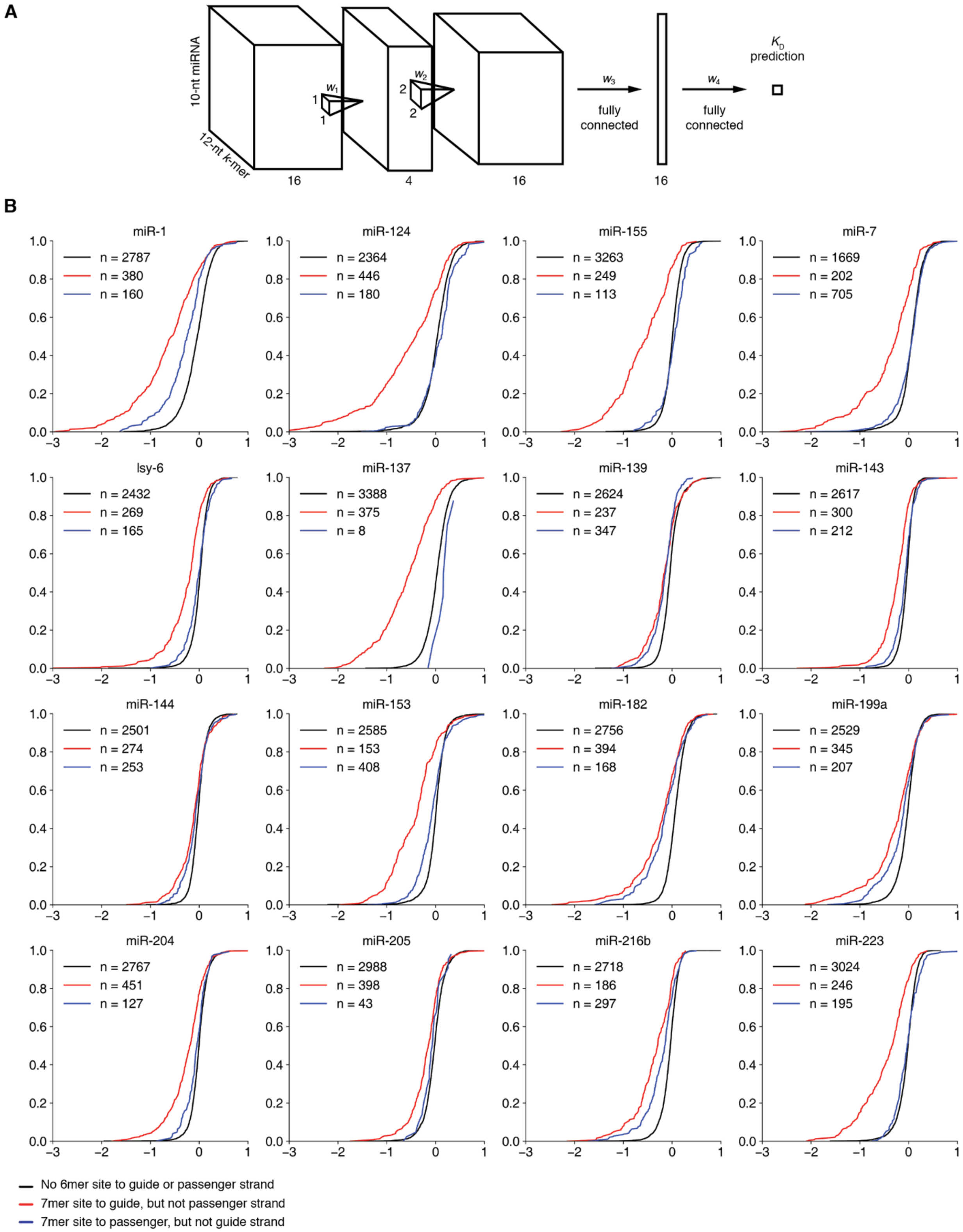


Fig. S9. Additional analyses and data related to training the CNN. (**A**) Schematic of the CNN architecture. Each miRNA and 12-nt *k*-mer pair was represented by a 10×12×16 matrix, where [*i*, *j*, 1 : 16] represented the one-hot encoding of the *i*th nucleotide of the miRNA and the

*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, which 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 7–8-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).

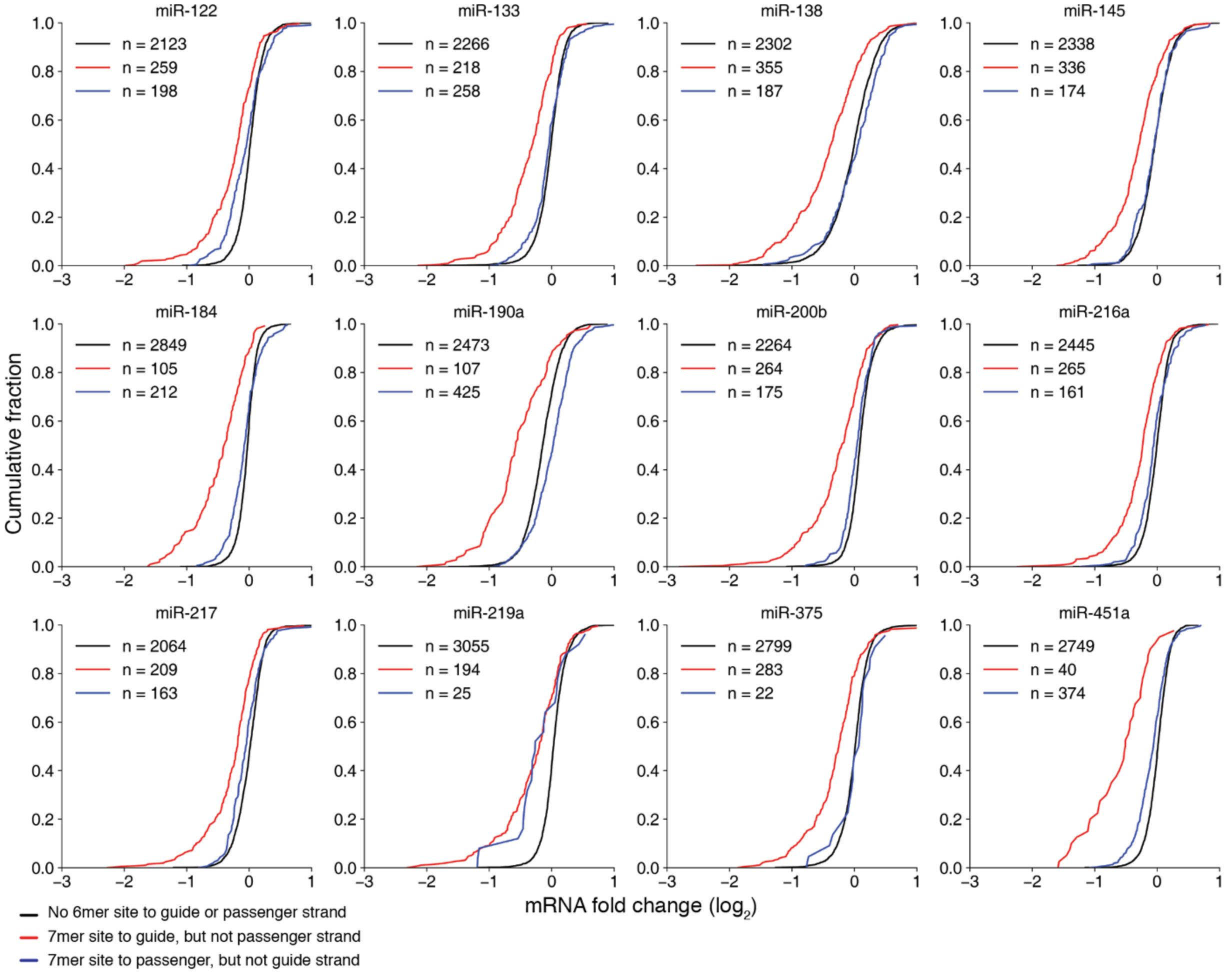


Fig. S10. Response of mRNAs to transfected miRNAs used for testing. Each plot shows cumulative distributions of fold-change values of mRNAs in HEK293FT cells. Otherwise, as in fig. S9B.

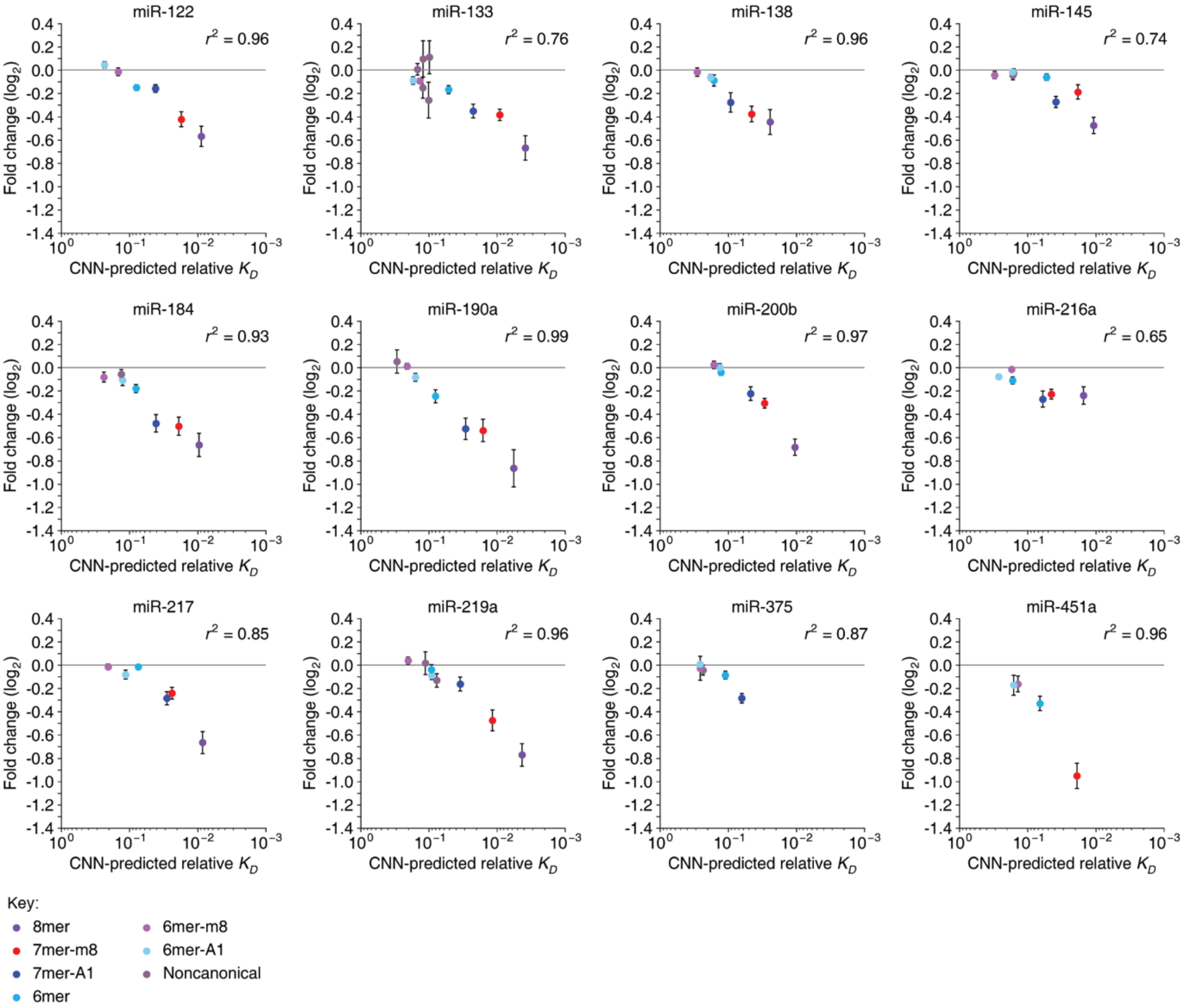


Fig. S11. Relationship between mean fold change conferred by each site type in HEK293FT cells and CNN-predicted relative K_D values. Results are shown for the six canonical site types and the predicted noncanonical sites found by examining the 12-nt *k*-mers that had the highest-affinity CNN-predicted K_D values but lacked a canonical site. Otherwise, as in Fig. 3, D to I.

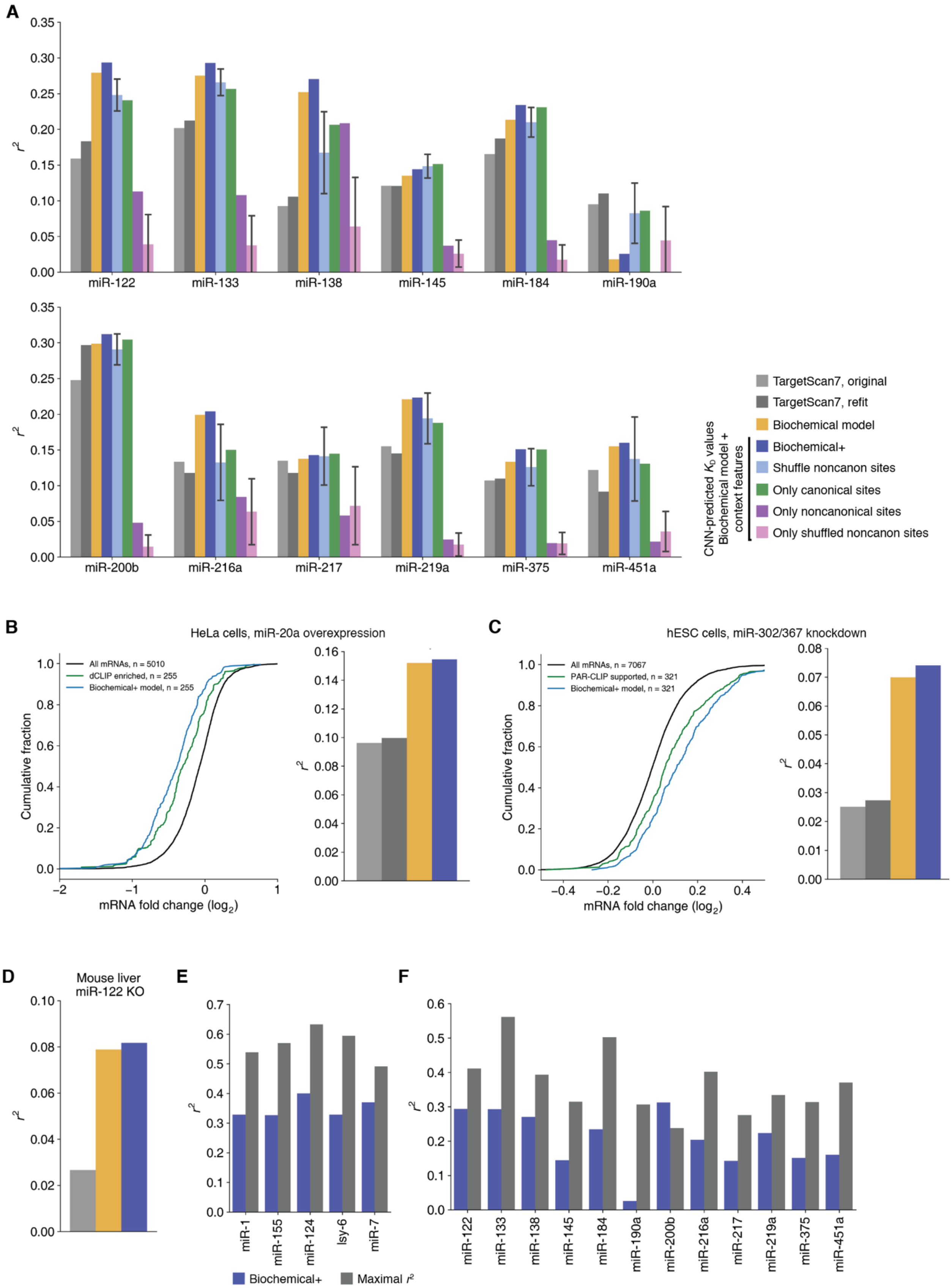

A
r²
miR-122
miR-133
miR-138
miR-145
miR-184
miR-190a
miR-200b
miR-216a
miR-217
miR-219a
miR-375
miR-451a
TargetScan7, original
TargetScan7, refit
Biochemical model
CNN-predicted K_D values
Biochemical model +
context features
Biochemical+
Shuffle noncanon sites
Only canonical sites
Only noncanonical sites
Only shuffled noncanon sites
B
HeLa cells, miR-20a overexpression
All mRNAs, n = 5010
dCLIP enriched, n = 255
Biochemical+ model, n = 255
Cumulative fraction
mRNA fold change ($\log_2$)
r²
C
hESC cells, miR-302/367 knockdown
All mRNAs, n = 7067
PAR-CLIP supported, n = 321
Biochemical+ model, n = 321
Cumulative fraction
mRNA fold change ($\log_2$)
r²
D
Mouse liver
miR-122 KO
r²
E
r²
miR-1
miR-155
miR-124
lsy-6
miR-7
Biochemical+
Maximal r²
F
r²
miR-122
miR-133
miR-138
miR-145
miR-184
miR-190a
miR-200b
miR-216a
miR-217
miR-219a
miR-375
miR-451a

Fig. 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 Fig. 6. Otherwise, as in Fig. 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 fig. S8, A and B. (**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 fig. S8A. (**E**) Performance of the biochemical+ model (blue) compared with estimated maximal r^2 values (grey) for each of the five miRNAs in Fig. 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 Fig. 6. Otherwise, as in (E).

Tables

Table S1. Coefficients of linear effects in model of miRNA, site, and flanking-dinucleotide sequence contribution to site binding affinity; related to Fig. 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 Fig. 4D.

	Δln(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

Acknowledgements

We thank K. Heindl, T. Eisen, and T. Bepler for helpful discussions, Y. Zhou for providing processed CLIP data from miR-20a overexpression experiments, and members of the Bartel lab for comments on this manuscript. 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.

References

Agarwal, Vikram, George W. Bell, Jin Wu Nam, and David P. Bartel. 2015. "Predicting Effective MicroRNA Target Sites in Mammalian MRNAs." *ELife* 4 (AUGUST2015). https://doi.org/10.7554/eLife.05005.

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." *Nature Biotechnology*. https://doi.org/10.1038/nbt.3300.

Bartel, David P. 2009. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell* 136 (2): 215–33. https://doi.org/10.1016/j.cell.2009.01.002.

Bartel, David P. 2018. "Metazoan MicroRNAs." *Cell*. https://doi.org/10.1016/j.cell.2018.03.006.

Becker, Winston R., Benjamin Ober-Reynolds, Karina Jouravleva, Samson M. Jolly, Phillip D. Zamore, and William J. Greenleaf. 2019. "High-Throughput Analysis Reveals Rules for Target RNA Binding and Cleavage by AGO2." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2019.06.012.

Bosson, Andrew D., Jesse R. Zamudio, and Phillip A. Sharp. 2014. "Endogenous MiRNA and Target Concentrations Determine Susceptibility to Potential CeRNA Competition." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2014.09.018.

Brennecke, Julius, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. 2005. "Principles of MicroRNA-Target Recognition." In *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0030085.

Chandradoss, Stanley D., Nicole T. Schirle, Malwina Szczepaniak, Ian J. Macrae, and Chirlmin Joo. 2015. "A Dynamic Search Process Underlies MicroRNA Targeting." *Cell*. https://doi.org/10.1016/j.cell.2015.06.032.

Chi, Sung Wook, Gregory J. Hannon, and Robert B. Darnell. 2012. "An Alternative Mode of MicroRNA Target Recognition." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/nsmb.2230.

Cuperus, Josh T., Benjamin Groves, Anna Kuchina, Alexander B. Rosenberg, Nebojsa Jojic, Stanley Fields, and Georg Seelig. 2017. "Deep Learning of the Regulatory Grammar of Yeast 5′ Untranslated Regions from 500,000 Random Sequences." *Genome Research*. https://doi.org/10.1101/gr.224964.117.

Denzler, Rémy, Vikram Agarwal, Joanna Stefano, David P. Bartel, and Markus Stoffel. 2014. "Assessing the CeRNA Hypothesis with Quantitative Measurements of MiRNA and Target Abundance." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2014.03.045.

Denzler, Rémy, Sean E. McGeary, Alexandra C. Title, Vikram Agarwal, David P. Bartel, and Markus Stoffel. 2016. "Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2016.09.027.

Dignam, John David, Russell M. Lebovitz, and Robert G. Roeder. 1983. "Accurate Transcription Initiation by RNA Polymerase II in a Soluble Extract from Isolated Mammalian Nuclei." *Nucleic Acids Research*. https://doi.org/10.1093/nar/11.5.1475.

Dominguez, Daniel, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, et al. 2018. "Sequence, Structure, and Context Preferences of Human RNA Binding Proteins." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2018.05.001.

Eichhorn, Stephen W., Huili Guo, Sean E. McGeary, Ricard A. Rodriguez-Mias, Chanseok Shin, Daehyun Baek, Shu hao Hsu, Kalpana Ghoshal, Judit Villén, and David P. Bartel. 2014. “MRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues.” *Molecular Cell* 56 (1): 104–15. https://doi.org/10.1016/j.molcel.2014.08.028.

Flores-Jasso, C. Fabián, William E. Salomon, and Phillip D. Zamore. 2013. “Rapid and Specific Purification of Argonaute-Small RNA Complexes from Crude Cell Lysates.” *RNA*. https://doi.org/10.1261/rna.036921.112.

Friedman, Robin C., Kyle Kai How Farh, Christopher B. Burge, and David P. Bartel. 2009. “Most Mammalian MRNAs Are Conserved Targets of MicroRNAs.” *Genome Research* 19 (1): 92–105. https://doi.org/10.1101/gr.082701.108.

Garcia, David M, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. 2011. “Weak Seed-Pairing Stability and High Target-Site Abundance Decrease the Proficiency of Lsy-6 and Other MicroRNAs.” *Nature Structural & Molecular Biology* 18 (10): 1139–46. https://doi.org/10.1038/nsmb.2115.

Grimson, Andrew, Kyle Kai How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. 2007. “MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.” *Molecular Cell* 27 (1): 91–105. https://doi.org/10.1016/j.molcel.2007.06.017.

Grosswendt, Stefanie, Andrei Filipchyk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, and Nikolaus Rajewsky. 2014. “Unambiguous Identification of MiRNA: Target Site Interactions by Different Types of Ligation Reactions.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2014.03.049.

Gumienny, Rafal, and Mihaela Zavolan. 2015. “Accurate Transcriptome-Wide Prediction of MicroRNA Targets and Small Interfering RNA off-Targets with MIRZA-G.” *Nucleic Acids Research* 43 (3): 1380–91. https://doi.org/10.1093/nar/gkv050.

Hafner, Markus, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, et al. 2010. “Transcriptome-Wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP.” *Cell*. https://doi.org/10.1016/j.cell.2010.03.009.

Hausser, Jean, Markus Landthaler, Lukasz Jaskiewicz, Dimos Gaidatzis, and Mihaela Zavolan. 2009. “Relative Contribution of Sequence and Structure Features to the MRNA Binding of Argonaute/EIF2C-MiRNA Complexes and the Degradation of MiRNA Targets.” *Genome Research*. https://doi.org/10.1101/gr.091181.109.

Helwak, Aleksandra, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. 2013. “Mapping the Human MiRNA Interactome by CLASH Reveals Frequent Noncanonical Binding.” *Cell*. https://doi.org/10.1016/j.cell.2013.03.043.

Jaganathan, Kishore, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, et al. 2019. “Predicting Splicing from Primary Sequence with Deep Learning.” *Cell*. https://doi.org/10.1016/j.cell.2018.12.015.

Jens, Marvin, and Nikolaus Rajewsky. 2015. “Competition between Target Sites of Regulators Shapes Post-Transcriptional Gene Regulation.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3853.

Jo, Myung Hyun, Soochul Shin, Seung Ryoung Jung, Eunji Kim, Ji Joon Song, and Sungchul Hohng. 2015. “Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2015.04.027.

Jonas, Stefanie, and Elisa Izaurralde. 2015. "Towards a Molecular Understanding of MicroRNA-Mediated Gene Silencing." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3965.

Khorshid, Mohsen, Jean Hausser, Mihaela Zavolan, and Erik Van Nimwegen. 2013. "A Biophysical MiRNA-MRNA Interaction Model Infers Canonical and Noncanonical Targets." *Nature Methods*. https://doi.org/10.1038/nmeth.2341.

Kim, Doyeon, You Me Sung, Jinman Park, Sukjun Kim, Jongkyu Kim, Junhee Park, Haeok Ha, Jung Yoon Bae, Sohui Kim, and Daehyun Baek. 2016. "General Rules for Functional MicroRNA Targeting." *Nature Genetics*. https://doi.org/10.1038/ng.3694.

Klum, Shannon M, Stanley D Chandradoss, Nicole T Schirle, Chirlmin Joo, and Ian J MacRae. 2018. "Helix-7 in Argonaute2 Shapes the MicroRNA Seed Region for Rapid Target Recognition." *The EMBO Journal*. https://doi.org/10.15252/embj.201796474.

Kozomara, Ana, Maria Birgaoanu, and Sam Griffiths-Jones. 2019. "MiRBase: From MicroRNA Sequences to Function." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1141.

Lambert, Nicole, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. 2014. "RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins." *Molecular Cell* 54 (5): 887–900. https://doi.org/10.1016/j.molcel.2014.04.016.

Landgraf, Pablo, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, et al. 2007. "A Mammalian MicroRNA Expression Atlas Based on Small RNA Library Sequencing." *Cell* 129 (7): 1401–14. https://doi.org/10.1016/j.cell.2007.04.040.

Lewis, Benjamin P., Christopher B. Burge, and David P. Bartel. 2005. "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets." *Cell* 120 (1): 15–20. https://doi.org/10.1016/j.cell.2004.12.035.

Linsley, Peter S., Janell Schelter, Julja Burchard, Miho Kibukawa, Melissa M. Martin, Steven R. Bartz, Jason M. Johnson, et al. 2007. "Transcripts Targeted by the MicroRNA-16 Family Cooperatively Regulate Cell Cycle Progression." *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.02005-06.

Lipchina, Inna, Yechiel Elkabetz, Markus Hafner, Robert Sheridan, Aleksandra Mihailovic, Thomas Tuschl, Chris Sander, Lorenz Studer, and Doron Betel. 2011. "Genome-Wide Identification of MicroRNA Targets in Human ES Cells Reveals a Role for MiR-302 in Modulating BMP Response." *Genes and Development*. https://doi.org/10.1101/gad.17221311.

Loeb, Gabriel B., Aly A. Khan, David Canner, Joseph B. Hiatt, Jay Shendure, Robert B. Darnell, Christina S. Leslie, and Alexander Y. Rudensky. 2012. "Transcriptome-Wide MiR-155 Binding Map Reveals Widespread Noncanonical MicroRNA Targeting." *Molecular Cell*. https://doi.org/10.1016/j.molcel.2012.10.002.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. "ViennaRNA Package 2.0." *Algorithms for Molecular Biology*. https://doi.org/10.1186/1748-7188-6-26.

McGeary, Sean E, Kathy S Lin, Charlie Y Shi, Namita Bisaria, and David P Bartel. 2018. "The Biochemical Basis of MicroRNA Targeting Efficacy." *BioRxiv*, January, 414763. https://doi.org/10.1101/414763.

Nam, Jin Wu, Olivia S. Rissland, David Koppstein, Cei Abreu-Goodger, Calvin H. Jan, Vikram Agarwal, Muhammed A. Yildirim, Antony Rodriguez, and David P. Bartel. 2014. "Global Analyses of the Effect of Different Cellular Contexts on MicroRNA Targeting." *Molecular*

Cell. https://doi.org/10.1016/j.molcel.2014.02.013.

Nielsen, Cydney B., Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman, and Christopher B. Burge. 2007. "Determinants of Targeting by Endogenous and Exogenous MicroRNAs and SiRNAs." *RNA* 13 (11): 1894–1910. https://doi.org/10.1261/rna.768207.

Paraskevopoulou, Maria D., Georgios Georgakilas, Nikos Kostoulas, Ioannis S. Vlachos, Thanasis Vergoulis, Martin Reczko, Christos Filippidis, Theodore Dalamagas, and A. G. Hatzigeorgiou. 2013. "DIANA-MicroT Web Server v5.0: Service Integration into MiRNA Functional Analysis Workflows." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt393.

R Foundation for Statistical Computing. 2018. *R: A Language and Environment for Statistical Computing. Http://Www.R-Project.Org/.*

Rio, Donald C. 2013. "Expression and Purification of Active Recombinant T7 RNA Polymerase from E. Coli." *Cold Spring Harbor Protocols*. https://doi.org/10.1101/pdb.prot078527.

Salomon, William E, Samson M Jolly, Melissa J Moore, Phillip D Zamore, William E Salomon, Samson M Jolly, Melissa J Moore, Phillip D Zamore, and Victor Serebrov. 2015. "Single-Molecule Imaging Reveals That Argonaute Reshapes the Binding Properties of Its Nucleic Acid Article Single-Molecule Imaging Reveals That Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides." *Cell*. https://doi.org/10.1016/j.cell.2015.06.029.

Schirle, Nicole T., Jessica Sheu-Gruttadauria, Stanley D. Chandradoss, Chirlmin Joo, and Ian J. MacRae. 2015. "Water-Mediated Recognition of T1-Adenosine Anchors Argonaute2 to MicroRNA Targets." *ELife*. https://doi.org/10.7554/eLife.07646.

Schirle, Nicole T., Jessica Sheu-Gruttadauria, and Ian J. MacRae. 2014. "Structural Basis for MicroRNA Targeting." *Science*. https://doi.org/10.1126/science.1258040.

Schmiedel, Jörn M., Sandy L. Klemm, Yannan Zheng, Apratim Sahay, Nils Blüthgen, Debora S. Marks, and Alexander Van Oudenaarden. 2015. "MicroRNA Control of Protein Expression Noise." *Science*. https://doi.org/10.1126/science.aaa1738.

Shin, Chanseok, Jin Wu Nam, Kyle Kai How Farh, H. Rosaria Chiang, Alena Shkumatava, and David P. Bartel. 2010. "Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing." *Molecular Cell*.

Tafer, Hakim, Stefan L. Ameres, Gregor Obernosterer, Christoph A. Gebeshuber, Renée Schroeder, Javier Martinez, and Ivo L. Hofacker. 2008. "The Impact of Target Site Accessibility on the Design of Effective SiRNAs." *Nature Biotechnology*. https://doi.org/10.1038/nbt1404.

Tunney, Robert, Nicholas J. McGlincy, Monica E. Graham, Nicki Naddaf, Lior Pachter, and Liana F. Lareau. 2018. "Accurate Design of Translational Output by a Neural Network Model of Ribosome Distribution." *Nature Structural and Molecular Biology*. https://doi.org/10.1038/s41594-018-0080-2.

Wee, Liang Meng, C. Fabián Flores-Jasso, William E. Salomon, and Phillip D. Zamore. 2012. "Argonaute Divides Its RNA Guide into Domains with Distinct Functions and RNA-Binding Properties." *Cell*. https://doi.org/10.1016/j.cell.2012.10.036.

Zhang, Kai, Xiaorong Zhang, Zhiqiang Cai, Jie Zhou, Ran Cao, Ya Zhao, Zonggui Chen, et al. 2018. "A Novel Class of MicroRNA-Recognition Elements That Function Only within Open Reading Frames." Nature Structural and Molecular Biology. https://doi.org/10.1038/s41594-018-0136-3.

Chapter 3. Future directions

Expanding the deep learning approach for predicting miRNA targets

Our work so far has shown that binding affinity is the major determinant of miRNA targeting efficacy (Chapter 2), and predictions of binding affinities between miRNAs and their targets will only improve with more binding-affinity measurements. However, even when the binding affinities are experimentally determined, the calculated occupancy of AGO–miRNA complexes on an mRNA cannot explain more than 60% of the reproducible variability in mRNA abundance changes induced by a miRNA. There must therefore be other factors that help determine the extent of miRNA targeting efficacy, including the structural accessibility of the site in the mRNA, the cellular localization of the mRNA, competition or enhancement from other RNA-binding proteins, and other yet unknown factors. We attempted to manually curate and estimate the effects of some of these factors, but the improvement to target prediction was modest (Chapter 2).

The most promising avenue for substantially improving miRNA target predictions may be an unbiased machine learning approach where the model can learn from entire mRNA sequences. Such an approach has been successful in other similar contexts, including predicting the translational output of yeast mRNAs from their 5′-UTR sequences (Cuperus et al. 2017) and predicting splice sites from pre-mRNA sequences (Jaganathan et al. 2019). These models have the opportunity to learn long-range effects and nonlinear interactions between features at the expense of reduced interpretability. Some of these features, such as pairing of a potential target site to 3′-end of the miRNA, may be miRNA specific and require input of the miRNA sequence, similar to the current CNN that predicts binding affinities from miRNA and target sequences (Chapter 2). Other features may be specific to the cell-type in which the training data is collected, such as those caused by interactions between AGO and other RNA-binding proteins with cell-type specific expression profiles. While these features would only be useful for predicting miRNA activity in those cell-types, they could elucidate previously unappreciated interactors with the miRNA targeting pathway that only interact with AGO in certain contexts.

Still other features could be global features of mRNAs that affect miRNA targeting, such as cooperativity between miRNA target sites. Previous work (Sætrom et al. 2007; Grimson et al. 2007), as well as additional analysis in Appendix 3, have shown that two target sites that are close together on an mRNA (but not so close that two AGO complexes cannot bind at the same time) confer more repression than two target sites that are further away from each other. While estimates on the exact range for observable cooperativity differ slightly, the effect seems to be robust. Because the exact mechanism causing the cooperative effect is unknown, there is currently no way to quantitatively estimate its effect from first principles. However, a data-driven approach could potentially learn the extent of cooperativity between two sites given the sequences of the two sites and the linker region.

Another global feature that affects miRNA targeting is the structural accessibility of target sites. While *in silico* predictions of RNA accessibility have been marginally useful in assisting the prediction of miRNA targets (Agarwal et al. 2015), they depend on nearest-neighbor rules derived from *in vitro* measurements of RNA pairing activity, which do not accurately reflect intracellular environments, especially in terms of salt concentrations (Becker et al. 2019). mRNA molecules in eukaryotic cells are also generally less structured than those observed *in vitro*, as determined by chemical probing (Rouskin et al. 2014; Guo and Bartel 2016), suggesting that factors in the cytoplasm may be preventing the folding of mRNAs or actively unfolding them. While this is expected to be due, in part, to ribosomes unfolding mRNA structures during translation, this disparity in mRNA accessibility between *in vivo* and *in vitro* has been observed to be similar between coding and untranslated regions (Rouskin et al. 2014).

Large-scale miRNA transfection datasets could therefore potentially be incredibly useful for developing a model to predict RNA accessibility *in vivo*. Dozens of miRNA duplexes with diverse sequences could be transfected, included non-endogenous ones, that could evenly cover the 3′-UTR sequences of endogenous mRNAs, and the CNN developed in Chapter 2 can be used to predict the expected binding affinity between any transfected miRNA and potential target site. The discrepancy between predicted occupancy and observed repression can then be used by a model to learn rules governing RNA accessibility from millions of examples of miRNA–target interactions, and these rules could be made more robust by training on data collected from many different cell-types. The modularity of the miRNA and AGO system, the finding that more AGO binding leads to more repression in a predictable way, and the existence of a global binding predictor for miRNAs are all key to this strategy. It is perhaps fitting that while RNA structure-prediction algorithms have long been used to improve predictions of miRNA targeting efficacy, miRNA occupancy predictions may in turn be able to improve predictions of RNA structural accessibility.

References

Agarwal, Vikram, George W. Bell, Jin Wu Nam, and David P. Bartel. 2015. "Predicting Effective MicroRNA Target Sites in Mammalian MRNAs." *ELife* 4 (AUGUST2015). https://doi.org/10.7554/eLife.05005.

Becker, Winston R, Inga Jarmoskaite, Kalli Kappel, Pavanapuresan P Vaidyanathan, Sarah K Denny, Rhiju Das, William J Greenleaf, and Daniel Herschlag. 2019. "Quantitative High-Throughput Tests of Ubiquitous RNA Secondary Structure Prediction Algorithms via RNA/Protein Binding." *BioRxiv*, January, 571588. https://doi.org/10.1101/571588.

Cuperus, Josh T., Benjamin Groves, Anna Kuchina, Alexander B. Rosenberg, Nebojsa Jojic, Stanley Fields, and Georg Seelig. 2017. "Deep Learning of the Regulatory Grammar of Yeast 5′ Untranslated Regions from 500,000 Random Sequences." *Genome Research*. https://doi.org/10.1101/gr.224964.117.

Grimson, Andrew, Kyle Kai How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. 2007. "MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing." *Molecular Cell* 27 (1): 91–105. https://doi.org/10.1016/j.molcel.2007.06.017.

Guo, Junjie U., and David P. Bartel. 2016. "RNA G-Quadruplexes Are Globally Unfolded in Eukaryotic Cells and Depleted in Bacteria." *Science*. https://doi.org/10.1126/science.aaf5371.

Jaganathan, Kishore, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, et al. 2019. "Predicting Splicing from Primary Sequence with Deep Learning." *Cell*. https://doi.org/10.1016/j.cell.2018.12.015.

Rouskin, Silvi, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S. Weissman. 2014. "Genome-Wide Probing of RNA Structure Reveals Active Unfolding of MRNA Structures in Vivo." *Nature*. https://doi.org/10.1038/nature12894.

Sætrom, Pål, Bret S.E. Heale, Ola Snøve, Lars Aagaard, Jessica Alluin, and John J. Rossi. 2007. "Distance Constraints between MicroRNA Target Sites Dictate Efficacy and Cooperativity." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm133.

Appendix 1. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution

Sahin Naqvi[1,2,4], Daniel W. Bellott[1,4], Kathy S. Lin[1,3] and David C. Page[1,2,4]

[1]Whitehead Institute, Cambridge, Massachusetts 02142, USA;
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
[3]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;
[4]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts 02142, USA

S.N., D.W.B., and D.C.P. designed the study. S.N. performed analyses with assistance from D.W.B. K.S.L. developed and implemented the step-detection algorithm. S.N. and D.C.P. wrote the manuscript.

Published as:
Naqvi S, Bellott DW, Lin KS, Page DC. “Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution.” 2018. *Genome Research* 28(4): 474–483.

Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution

Sahin Naqvi,[1,2] Daniel W. Bellott,[1] Kathy S. Lin,[1,3] and David C. Page[1,2,4]

[1]*Whitehead Institute, Cambridge, Massachusetts 02142, USA;* [2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [3]*Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [4]*Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts 02142, USA*

Mammalian X and Y Chromosomes evolved from an ordinary autosomal pair. Genetic decay of the Y led to X Chromosome inactivation (XCI) in females, but some Y-linked genes were retained during the course of sex chromosome evolution, and many X-linked genes did not become subject to XCI. We reconstructed gene-by-gene dosage sensitivities on the ancestral autosomes through phylogenetic analysis of microRNA (miRNA) target sites and compared these preexisting characteristics to the current status of Y-linked and X-linked genes in mammals. Preexisting heterogeneities in dosage sensitivity, manifesting as differences in the extent of miRNA-mediated repression, predicted either the retention of a Y homolog or the acquisition of XCI following Y gene decay. Analogous heterogeneities among avian Z-linked genes predicted either the retention of a W homolog or gene-specific dosage compensation following W gene decay. Genome-wide analyses of human copy number variation indicate that these heterogeneities consisted of sensitivity to both increases and decreases in dosage. We propose a model of XY/ZW evolution incorporating such preexisting dosage sensitivities in determining the evolutionary fates of individual genes. Our findings thus provide a more complete view of the role of dosage sensitivity in shaping the mammalian and avian sex chromosomes and reveal an important role for post-transcriptional regulatory sequences (miRNA target sites) in sex chromosome evolution.

[Supplemental material is available for this article.]

The mammalian X and Y Chromosomes evolved from a pair of ordinary autosomes over the past 300 myr (Lahn and Page 1999). Only 3% of genes on the ancestral pair of autosomes survive on the human Y Chromosome (Skaletsky et al. 2003; Bellott et al. 2010) compared with 98% on the X Chromosome (Mueller et al. 2013). In females, one copy of the X Chromosome is silenced by X-inactivation (XCI); this silencing evolved on a gene-by-gene basis following Y gene loss in males and X up-regulation in both sexes (Jegalian and Page 1998; Ross et al. 2005; Berletch et al. 2015; Tukiainen et al. 2017), and some genes escape XCI in humans (Carrel and Willard 2005) and other mammals (Yang et al. 2010). Dosage compensation refers to any mechanism restoring ancestral dosage following gene loss from the sex-specific chromosome. In mammalian males, therefore, dosage compensation consisted solely of X up-regulation, as it returned X-linked gene expression to ancestral levels following Y gene loss. In females, dosage compensation involved both X up-regulation and the acquisition of XCI, which increased and decreased X-linked expression levels, respectively. Since females did not undergo any initial decrease in ancestral dosage due to Y gene loss, X up-regulation and the acquisition of XCI together restored ancestral expression levels.

In parallel, the avian Z and W sex chromosomes evolved from a different pair of autosomes than the mammalian X and Y Chromosomes (Nanda et al. 1999; Ross et al. 2005; Bellott et al. 2010). Decay of the female-specific W Chromosome was similarly extensive, but birds did not evolve a large-scale inactivation of Z-linked genes analogous to XCI in mammals (Itoh et al. 2007). Dosage compensation, as measured by a male/female expression ratio close to one, has been observed for some Z-linked genes in some tissues (Mank and Ellegren 2009; Uebbing et al. 2015; Zimmer et al. 2016). Thus, genes previously found on the ancestral autosomes that gave rise to the mammalian or avian sex chromosomes have undergone significant changes in gene dosage. In modern mammals, these molecular events have resulted in three classes of ancestral X-linked genes representing distinct evolutionary fates: those with a surviving Y homolog, those with no Y homolog and subject to XCI, and those with no Y homolog but escaping XCI. In birds, two clear classes of ancestral Z-linked genes have arisen: those with or without a W homolog, with additional heterogeneity among Z-linked genes without a W homolog as a result of gene-specific dosage compensation. Identifying gene-by-gene properties that distinguish classes of X- and Z-linked genes is thus crucial to understanding the selective pressures underlying the molecular events of mammalian and avian sex chromosome evolution.

Emerging evidence suggests a role for gene dosage sensitivity in mammalian and avian sex chromosome evolution. X- and Z-linked genes with surviving homologs on the mammalian Y or avian W Chromosomes are enriched for important regulatory functions and predictors of haploinsufficiency compared with those lacking Y or W homologs (Bellott et al. 2014, 2017); similar

Corresponding author: dcpage@wi.mit.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.230433.117. Freely available online through the *Genome Research* Open Access option.

observations have been made in fish (White et al. 2015) and *Drosophila* (Kaiser et al. 2011). Human X- and chicken Z-linked genes that show the strongest signatures of dosage compensation in either lineage also show signs of dosage sensitivity as measured by membership in large protein complexes (Pessia et al. 2012) or evolutionary patterns of gene duplication and retention (Zimmer et al. 2016). Despite these advances, little is known regarding selective pressures resulting from sensitivity to dosage increases, as these studies either focused on haploinsufficiency or employed less direct predictors of dosage sensitivity. Furthermore, it is not known whether heterogeneities in dosage sensitivity among classes of sex-linked genes were acquired during sex chromosome evolution, or predated the emergence of sex chromosomes, as there has been no explicit, systematic reconstruction of dosage sensitivity on the ancestral autosomes that gave rise to the mammalian and avian sex chromosomes.

To assess the role of preexisting dosage sensitivities in XY and ZW evolution, we sought to employ a measure of dosage sensitivity that could be (1) demonstrably informative with respect to sensitivity to dosage increases and (2) explicitly reconstructed on the ancestral autosomes. We focused on regulation by microRNAs (miRNAs), small noncoding RNAs that function as tuners of gene dosage by lowering target mRNA levels through pairing to the 3′ untranslated region (UTR) (Bartel 2009). The repressive nature of miRNA targeting is informative with respect to sensitivity to dosage increases, allowing for a more complete understanding of the role of dosage sensitivity in sex chromosome evolution. Both miRNAs themselves and their complementary target sites can be preserved over millions of years of vertebrate evolution, facilitating the reconstruction of miRNA targeting on the ancestral autosomes through cross-species sequence alignments. As miRNA targeting occurs post-transcriptionally, reconstruction of its ancestral state is decoupled from transcriptional regulatory mechanisms such as XCI that evolved following X-Y differentiation.

Results

Analysis of human copy number variation indicates conserved miRNA targeting of genes sensitive to dosage increases

We first sought to determine whether conserved targeting by miRNAs correlates with sensitivity to dosage increases across the human genome. To estimate pressure to maintain miRNA targeting, we used published probabilities of conserved targeting (P_{CT} scores) for each gene–miRNA interaction in the human genome. The P_{CT} score reflects an estimate of the probability that a given gene–miRNA interaction is conserved due to miRNA targeting, obtained by calculating the conservation of the relevant miRNA target sites relative to the conservation of the entire 3′ UTR (Friedman et al. 2009). In this manner, the P_{CT} score intrinsically controls for differences in background conservation and sequence composition, both of which vary widely among 3′ UTRs due to differing rates of expression divergence and/or sequence evolution. We refer to these P_{CT} scores as "miRNA conservation scores" in the remainder of the text.

A recent study reported a correlation between these miRNA conservation scores and predicted haploinsufficiency (Pinzón et al. 2017), indicating that conserved miRNA targeting broadly corresponds to dosage sensitivity. However, such a correlation does not isolate the effects of sensitivity to dosage increases, which we expect to be particularly important in the context of miRNA targeting. We reasoned that genes for which increases in dosage are deleterious should be depleted from the set of observed gene duplications in healthy human individuals. We used a catalog of rare genic copy number variation among 59,898 control human exomes (Exome Aggregation Consortium [ExAC]) (Ruderfer et al. 2016) to classify autosomal protein-coding genes as exhibiting or lacking duplication or deletion in healthy individuals (see Methods). We compared duplicated and nonduplicated genes with the same deletion status in order to control for differences in sensitivity to underexpression. We found that nonduplicated genes have significantly higher miRNA conservation scores than duplicated genes, irrespective of deletion status (Fig. 1A,B). Nondeleted genes also have significantly higher scores than deleted genes irrespective of duplication status (Supplemental Fig. S1), but duplication status has a greater effect on miRNA conservation scores than does deletion status (Fig. 1C, blue vs. orange boxes). Thus, conserved miRNA targeting is a feature of genes sensitive to changes in gene dosage in humans and is especially informative with regards to sensitivity to dosage increases, consistent with the known role of miRNAs in tuning gene dosage by lowering target mRNA levels.

X-Y pairs and X-inactivated genes have higher miRNA conservation scores than X escape genes

We next assessed whether the three classes of X-linked genes differ with respect to dosage sensitivity as inferred by conserved miRNA targeting. To delineate these classes, we began with the set of ancestral genes reconstructed through cross-species comparisons of the human X Chromosome and orthologous chicken autosomes (Bellott et al. 2010, 2014, 2017; Hughes et al. 2012; Mueller et al. 2013). We designated ancestral X-linked genes with a surviving human Y homolog (Skaletsky et al. 2003) as X-Y pairs and also considered the set of X-linked genes with a surviving Y homolog in any of eight mammals (Bellott et al. 2014) to increase the

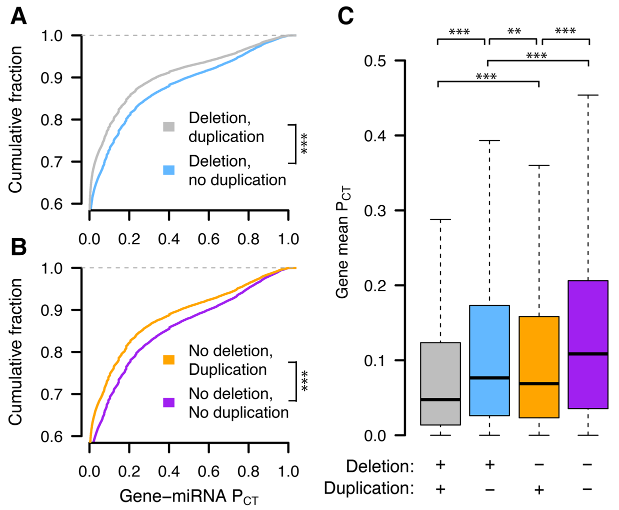


Figure 1. Conserved miRNA targeting of autosomal genes stratified by copy number variation in 59,898 human exomes. Probabilities of conserved targeting (P_{CT}) of all gene–miRNA interactions involving nonduplicated and duplicated genes, further stratified as (*A*) deleted (gray, n = 69,339 interactions from 4118 genes; blue, n = 80,290 interactions from 3976 genes) or (*B*) not deleted (orange, n = 51,514 interactions from 2916 genes; purple, n = 72,826 interactions from 3510 genes). (***) P < 0.001, two-sided Kolmogorov–Smirnov test. (*C*) Mean gene-level P_{CT} scores. (**) P < 0.01, (***) P < 0.001, two-sided Wilcoxon rank-sum test.

phylogenetic breadth of findings regarding X-Y pairs. A number of studies have cataloged the inactivation status of X-linked genes in various human tissues and cell types. We used a meta-analysis that combined results from three studies by assigning a "consensus" X-inactivation status to each gene (Balaton et al. 2015) to designate the remainder of ancestral genes lacking a Y homolog as subject to or escaping XCI. In summary, we classified genes as either: (1) X-Y pairs, (2) lacking a Y homolog and subject to XCI (X-inactivated), or (3) lacking a Y homolog but escaping XCI (X escape).

We found that human X-Y pairs have the highest miRNA conservation scores, followed by X-inactivated and finally X escape genes (Fig. 2A,B). The expanded set of X-Y pairs across eight mammals also has significantly higher miRNA conservation scores than ancestral X-linked genes with no Y homolog (Supplemental Fig. S2). Observed differences between miRNA conservation scores are not driven by distinct subsets of genes in each class, as indicated by gene resampling with replacement (Supplemental Fig. S3). The decrease in miRNA conservation scores of X escape genes relative to X-inactivated genes and X-Y pairs is not driven by genes that escape XCI variably across individuals (Supplemental Fig. S4), and was consistent even when including ambiguous genes as either X-inactivated or X escape genes (Supplemental Fig. S5). We also verified that these differences were not driven by artificially inflated or deflated conservation scores of certain target sites due to nonuniformity in 3′ UTR conservation (Methods; Supplemental Fig. S6).

Finally, we assessed whether miRNA conservation scores distinguish the three classes by providing additional information not accounted for by known factors (Bellott et al. 2014) influencing evolutionary outcomes. We used logistic regression to model, for each gene, the probability of falling into each of the three classes (X-Y pair, X-inactivated, or X escape) as a linear combination of haploinsufficiency probability (pHI) (Huang et al. 2010); human expression breadth (The GTEx Consortium 2015); purifying selection, measured by the ratio of nonsynonymous to synonymous substitution rates (d_N/d_S) between human and mouse orthologs (Yates et al. 2016); and mean gene-level miRNA conservation scores. We note that pHI is a score composed of several genic features, one of which is the number of protein–protein interactions, consistent with the idea that members of large protein complexes tend to be dosage-sensitive (Papp et al. 2003; Pessia et al. 2012). Removing either miRNA conservation or pHI as predictors from the full model resulted in inferior model fits as measured by the Akaike information criterion (AIC) (full model, AIC 321.5; full model minus miRNA conservation, AIC 327.7; full model minus pHI, AIC 327.3; higher AIC indicates inferior model). Therefore, miRNA conservation and pHI contribute independent information that distinguishes the three classes of X-linked genes. Based on our analyses of autosomal copy number variation (Fig. 1), we attribute this independence to the fact that miRNA conservation scores are most informative with regards to sensitivity to dosage increases. Taken together, these results indicate significant heterogeneity in dosage sensitivity, as inferred by miRNA target site conservation, among the three classes of ancestral X-linked genes: X-Y pairs are the most dosage-sensitive, while X-inactivated genes are of intermediate dosage sensitivity, and X escape genes are the least dosage-sensitive.

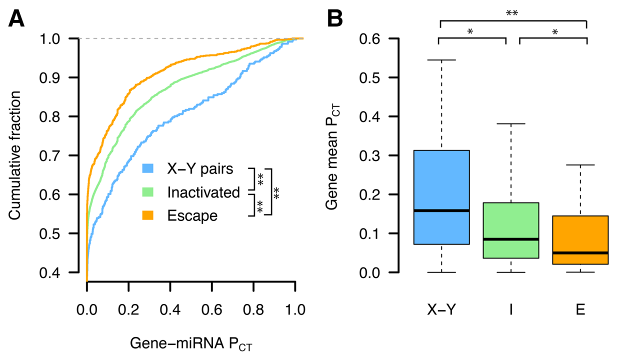


Figure 2. X-Y pairs and X-inactivated genes have higher miRNA conservation scores than X escape genes. P_{CT} score distributions of all gene–miRNA interactions involving (*A*) human X-Y pairs ($n = 371$ interactions from 15 genes), X-inactivated genes ($n = 6743$ interactions from 329 genes), and X escape genes ($n = 1037$ interactions from 56 genes). (**) $P < 0.01$, two-sided Kolmogorov–Smirnov test. (*B*) Mean gene-level P_{CT} scores. (*) $P < 0.05$, (**) $P < 0.01$, two-sided Wilcoxon rank-sum test.

Heterogeneities in X-linked miRNA targeting were present on the ancestral autosomes

We next asked whether differences in miRNA targeting were present on the ancestral autosomes that gave rise to the mammalian X and Y Chromosomes. To reconstruct the ancestral state of miRNA targeting, we first focused on miRNA target sites in the 3′ UTR of human orthologs that align with perfect identity to a site in the corresponding chicken ortholog; these sites were likely present in the common ancestor of mammals and birds (Fig. 3A). We found that X-Y pairs have the most human–chicken conserved target sites, followed by X-inactivated genes, and then X escape genes (Fig. 3B, top). Unlike the miRNA conservation scores used earlier, this metric does not account for background conservation; we therefore estimated the background conservation of each 3′ UTR using shuffled miRNA family seed sequences (see Methods). X-Y pairs, X-inactivated genes, and X escape genes do differ significantly with respect to background conservation (Supplementary Fig. S7), but these differences cannot account for the observed differences in true human–chicken conserved sites (Fig. 3B, bottom). We observed similar results for the expanded set of X-Y pairs across eight mammals (Supplemental Fig. S8A).

Differences in the number of human–chicken conserved sites among the three classes of X-linked genes could be explained by heterogeneity in miRNA targeting present on the ancestral autosomes or by ancestral homogeneity followed by different rates of target site loss during or following X-Y differentiation. To distinguish between these two possibilities, we took advantage of previous reconstructions of human sex chromosome evolution (Fig. 3A; Bellott et al. 2014), which confirmed that, following the divergence of placental mammals from marsupials, an X-autosome chromosomal fusion generated the X-added region (XAR) (Watson et al. 1990). Genes on the XAR are therefore X-linked in placental mammals but autosomal in marsupials such as the opossum. We limited our analysis to genes in the XAR and target sites conserved between orthologous chicken and opossum 3′ UTRs, ignoring site conservation in humans; these sites were likely present in the common ancestor of mammals and birds, and an absence of such sites cannot be explained by site loss following X-Y differentiation. We observed the same pattern as with the human–chicken conserved sites, both before and after accounting for background 3′ UTR conservation (Fig. 3C, three gene classes; Supplemental Fig. S8B, X-Y pairs across eight mammals). These results demonstrate that the autosomal precursors of X-Y pairs and X-inactivated genes were subject to more miRNA-mediated regulation than X

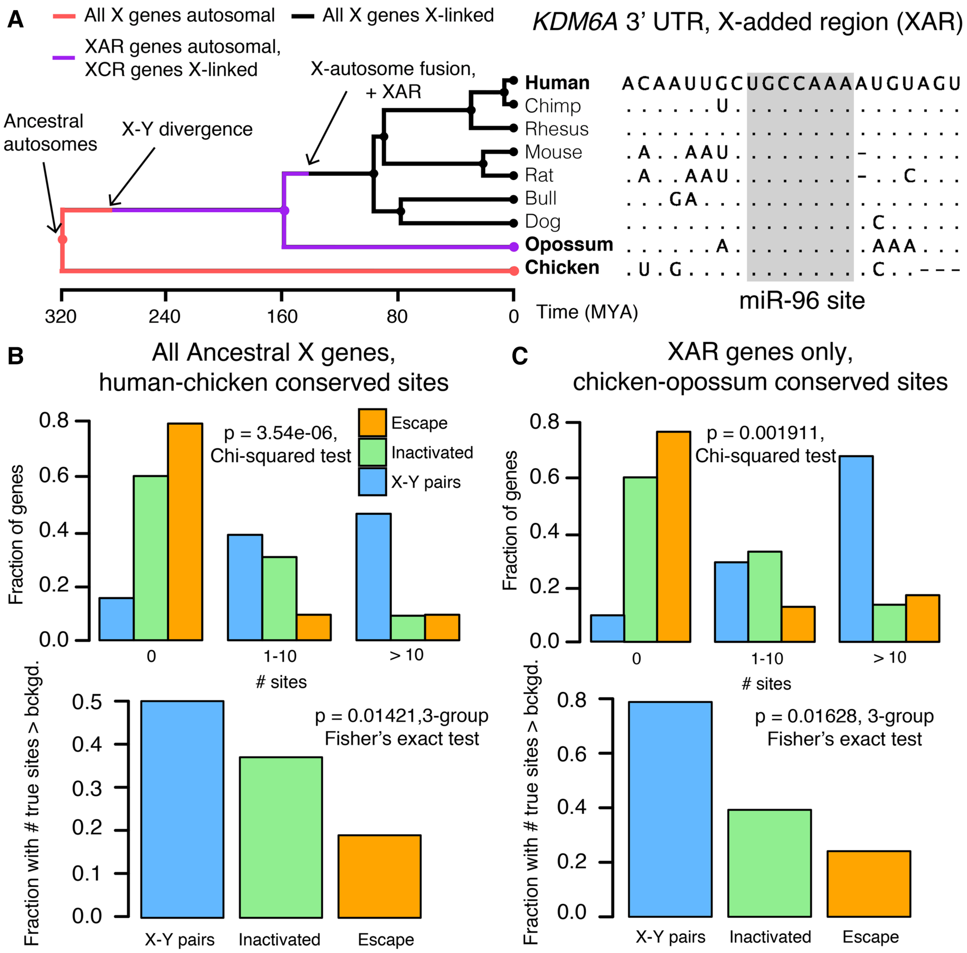


Figure 3. Heterogeneities in X-linked miRNA targeting were present on the ancestral autosomes. (*A*) Example reconstruction of an ancestral miR-96 target site in the 3′ UTR of *KDM6A*, an X-linked gene in the X-added region (XAR) with a surviving Y homolog. (*Left*) Species tree overlaid with events in mammalian sex chromosome evolution. (*Right*) Multiple sequence alignment; dots in nonhuman species indicate identity with human sequence; dashes indicate gaps in alignment. (*B*) Distributions of sites conserved between 3′ UTRs of human and chicken orthologs (*top*) or comparisons to background expectation (*bottom*; see Methods) for human X-Y pairs ($n = 16$), X-inactivated genes ($n = 251$), and X escape genes ($n = 42$). (*C*) Statistics as in *B*, but using sites conserved between chicken and opossum 3′ UTRs only for genes in the XAR: X-Y pairs ($n = 11$), X-inactivated genes ($n = 58$), and X escape genes ($n = 27$).

escape genes. Combined with our earlier results, we conclude that present-day heterogeneities in dosage sensitivity on the mammalian X Chromosome existed on the ancestral autosomes from which it derived.

Z-W pairs have higher miRNA conservation scores than other ancestral Z-linked genes

We next assessed whether classes of avian Z-linked genes, those with and without a W homolog, show analogous heterogeneities in sensitivity to dosage increases. We used the set of ancestral genes reconstructed through cross-species comparisons of the avian Z Chromosome and orthologous human autosomes and focused on the set of Z-W pairs identified by sequencing of the chicken W Chromosome (Bellott et al. 2010, 2017). To increase the phylogenetic breadth of our comparisons, we also included candidate Z-W pairs obtained through comparisons of male and female genome assemblies (four-species set) or inferred by read-depth changes in female genome assemblies (14-species set; for details, see Methods) (Zhou et al. 2014). The more complete 3′ UTR annotations in the human genome relative to chicken allow for a more accurate assessment of conserved miRNA targeting. Accordingly, we analyzed the 3′ UTRs of the human orthologs of chicken Z-linked genes.

We found that the human orthologs of Z-W pairs have higher miRNA conservation scores than the human orthologs of other ancestral Z genes (Fig. 4A,B). Differences in miRNA conservation scores between Z-W pairs and other ancestral Z genes remained significant when considering the expanded sets of Z-W pairs across four and 14 avian species (Supplemental Fig. S9). These differences are not driven by distinct subsets of genes, as indicated by gene resampling with a replacement (Supplemental Fig. S10), and cannot be accounted for by within-UTR variation in regional conservation (Supplemental Fig. S11). Logistic regression models indicate that miRNA conservation scores provide additional information not captured by known factors (Bellott et al. 2017) influencing survival of W-linked genes (full model, AIC 127.1; full model minus miRNA conservation, AIC 137.8; full model minus pHI, AIC 132.7; higher AIC indicates inferior model). Together, these results indicate that Z-linked genes with a surviving W homolog are more sensitive to changes in dosage—both increases and decreases—than are genes without a surviving W homolog.

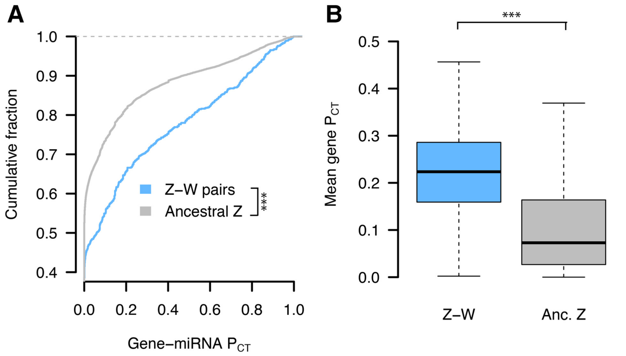


Figure 4. Z-W pairs have higher miRNA conservation scores than other ancestral Z-linked genes. (*A*) P$_{CT}$ score distributions of all gene–miRNA interactions involving the human orthologs of chicken Z-W pairs ($n = 832$ interactions from 28 genes) and other ancestral Z genes ($n = 16{,}692$ interactions from 657 genes). (***) $P < 0.001$, two-sided Kolmogorov–Smirnov test. (*B*) Mean gene-level P$_{CT}$ scores. (***) $P < 0.001$, two-sided Wilcoxon rank-sum test.

While there are two clear classes of Z-linked genes—those with or without a W homolog—studies of Z-linked gene expression have suggested additional heterogeneity among Z-linked genes without a W homolog due to gene-specific dosage compensation (Mank and Ellegren 2009; Uebbing et al. 2015; Zimmer et al. 2016). If Z-linked genes with no W homolog exist upon a continuum from noncompensated to dosage compensated, those that are more compensated should have more conserved miRNA target sites, reflective of greater dosage sensitivity. We quantified the dosage compensation by using RNA sequencing data (Marin et al. 2017) to compare, in four somatic tissues, the chicken male/female expression ratio to the analogous ratio in human and *Anolis* (see Methods). In the brain, kidney, and liver, Z-linked genes with no W homolog and higher mean miRNA conservation scores had male/female expression ratios closer to one (Supplemental Fig. S12). Thus, in addition to the above-described differences between Z-linked genes with or without a W homolog, Z-linked genes with no W homolog but with more effective dosage compensation have more conserved miRNA target sites than noncompensated genes.

Heterogeneities in Z-linked miRNA targeting were present on the ancestral autosomes

We next asked whether differences in miRNA targeting between Z-W pairs and other ancestral Z-linked genes were present on the ancestral autosomes that gave rise to the avian Z and W Chromosomes. We found that chicken Z-W pairs have more human–chicken-conserved miRNA target sites than their Z-linked counterparts without surviving W homologs, both before (Fig. 5B, top) and after (Fig. 5B, bottom) accounting for the background conservation of each individual 3′ UTR. To confirm that these differences represent ancestral heterogeneity rather than differential site loss during or following Z-W differentiation, we instead considered the number of sites conserved between human and *Anolis* lizard, which diverged from birds prior to Z-W differentiation (Fig. 5A). Chicken Z-W pairs contain an excess of human–*Anolis* conserved miRNA target sites, both before (Fig. 5C, top) and after (Fig. 5C, bottom) accounting for the background conservation of each individual 3′ UTR. We observed similar results with the predicted four-species (Supplemental Fig. S13) and 14-species (Supplemental Fig. S14) sets of Z-W pairs. Thus, the autosomal precursors of avian Z-W pairs were subject to more miRNA-mediated regulation than the autosomal precursors of Z-linked genes that lack a W homolog. Furthermore, in the liver and brain, Z-linked genes with no W homolog with an excess of human–chicken-conserved miRNA sites had male/female expression ratios closer to one, implying more effective dosage compensation (Supplemental Fig. S15). Together, these results indicate heterogeneity in dosage sensitivity among genes on the ancestral autosomes that gave rise to the avian Z Chromosome.

Analyses of experimental data sets validate miRNA target site function

Our results to this point, which indicate preexisting heterogeneities in dosage constraints among X- or Z-linked genes as inferred by predicted miRNA target sites, lead to predictions regarding the function of these sites in vivo. To test these predictions, we turned to publicly available experimental data sets consisting both of gene expression profiling following transfection or knockout of individual miRNAs, and of high-throughput crosslinking-immunoprecipitation (CLIP) to identify sites that bind Argonaute in vivo (see Methods). If the above-studied sites are effective in mediating target repression, targets of an individual miRNA should show increased expression levels or Argonaute binding following miRNA transfection and show decreased expression levels following miRNA knockout. Together, our analyses of publicly available data sets fulfilled these predictions, validating the function of these sites in multiple cellular contexts and species (Fig. 6). From the gene expression profiling data, we observed results consistent with effective targeting by (1) 11 different miRNA families in human HeLa cells (Supplemental Fig. S16), (2) four different miRNAs in human HCT116 and HEK293 cells (Supplemental Fig. S17), and (3) miR-155 in mouse B and Th1 cells (Supplemental Fig. S18). In the CLIP data, the human orthologs of X- or Z-linked targets of miR-124 are enriched for Argonaute-bound clusters that appear following miR-124 transfection, while a similar but nonsignificant enrichment is observed for miR-7 (Supplemental Fig. S19). Thus, conserved miRNA target sites used to infer dosage constraints on X-linked genes, and the autosomal orthologs of Z-linked genes can effectively mediate target repression in living cells.

Discussion

Here, through the evolutionary reconstruction of miRNA target sites, we provide evidence for preexisting heterogeneities in dosage sensitivity among genes on the mammalian X and avian Z Chromosomes. We first showed that, across all human autosomal genes, dosage sensitivity—as indicated by patterns of genic copy number variation—correlates with the degree of conserved miRNA targeting. We found that conserved targeting correlates especially strongly with sensitivity to dosage increases, consistent with miRNA targeting serving to reduce gene expression. Turning to the sex chromosomes of mammals and birds, genes that retained a homolog on the sex-specific Y or W Chromosome (X-Y and Z-W pairs) have more conserved miRNA target sites than genes with no Y or W homolog. In mammals, genes with no Y homolog that became subject to XCI have more conserved sites than those that continued to escape XCI following Y gene decay. In birds, across Z-linked genes with no W homolog, the degree of conserved miRNA targeting correlates with the degree of gene-specific dosage compensation. We then reconstructed the ancestral state of miRNA targeting, observing significant

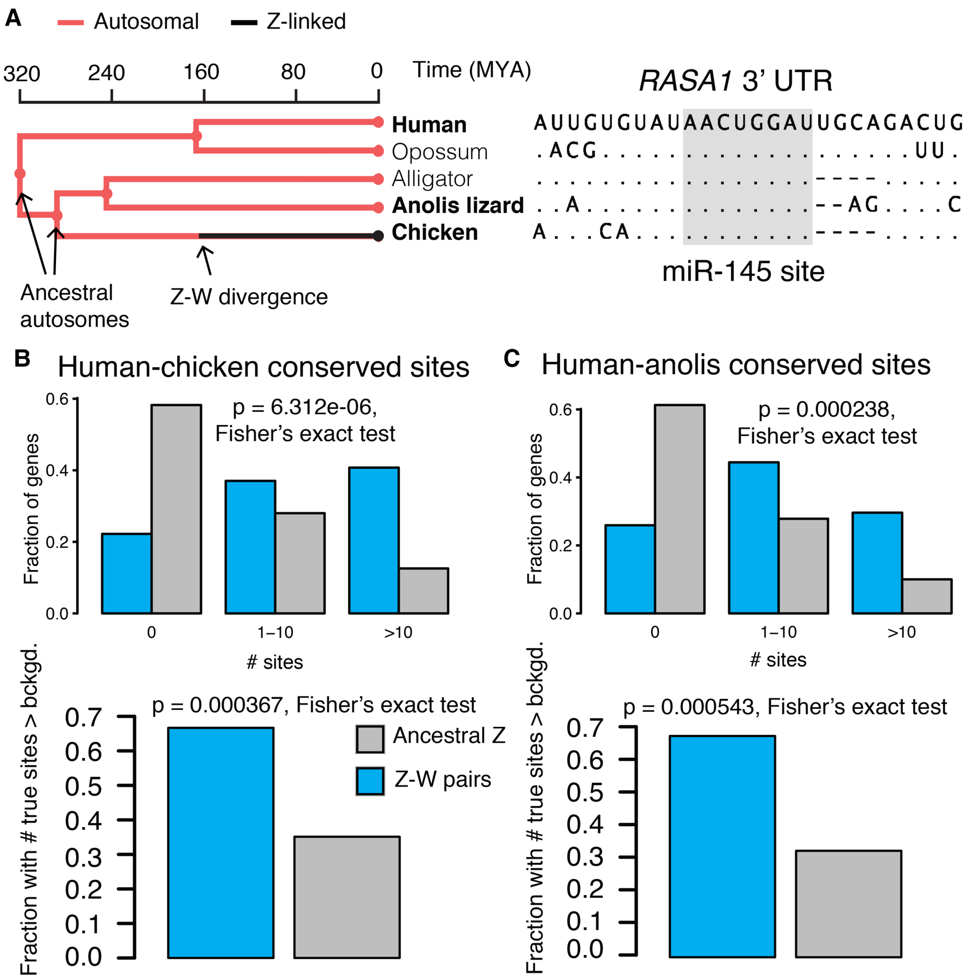


Figure 5. Heterogeneities in Z-linked miRNA targeting were present on the ancestral autosomes. (*A*) Example reconstruction of an ancestral miR-145 target site in the 3′ UTR of *RASA1*, a Z-linked gene with a surviving W homolog. (*Left*) Species tree overlaid with events in avian sex chromosome evolution. (*Right*) Multiple sequence alignment; dots in nonhuman species indicate identity with human sequence; dashes indicate gaps in alignment. (*B*) Numbers of sites conserved between 3′ UTRs of human and chicken orthologs (*top*) or comparisons to background expectation (*bottom*) for chicken Z-W pairs ($n = 27$) and other ancestral Z genes ($n = 578$). (*C*) Statistics as in *B*, but using sites conserved between human and *Anolis* 3′ UTRs.

heterogeneities in the extent of miRNA targeting, and thus dosage sensitivity, on the ancestral autosomes that gave rise to the mammalian and avian sex chromosomes. Finally, through analysis of publicly available experimental data sets, we validated the function, in living cells, of the miRNA target sites used to infer dosage sensitivity. We thus conclude that differences in dosage sensitivity —both to increases and to decreases in gene dosage—among genes on the ancestral autosomes influenced their evolutionary trajectory during sex chromosome evolution, not only on the sex-specific Y and W Chromosomes, but also on the sex-shared X and Z Chromosomes.

Our findings build upon previous work in three important ways. First, our analysis of miRNA-mediated repression indicates that these heterogeneities consist of sensitivities to dosage increases and decreases, whereas previous studies had either focused on sensitivity to underexpression or could not differentiate the two. Second, our reconstruction of miRNA targeting on the ancestral autosomes provides direct evidence that heterogeneities in dosage sensitivity among classes of X- and Z-linked were preexisting rather than acquired during sex chromosome evolution. Finally, by pointing to specific regulatory sequences (miRNA target sites) functioning to tune gene dosage both prior to and during sex chromosome evolution, our study provides a view of dosage compensation encompassing post-transcriptional regulation.

Human disease studies support the claim that increased dosage of X-Y pairs and X-inactivated genes is deleterious to fitness. Copy number gains of the X-linked gene *KDM6A*, which has a surviving human Y homolog, are found in patients with developmental abnormalities and intellectual disability (Lindgren et al. 2013). *HDAC6*, *CACNA1F*, *GDI1*, and *IRS4* all lack Y homologs and are subject to XCI in humans. A mutation in the 3′ UTR of *HDAC6* abolishing targeting by miR-433 has been linked to familial chondrodysplasia in both sexes (Simon et al. 2010). Likely gain-of-function mutations in *CACNA1F* cause congenital stationary night blindness in both sexes (Hemara-Wahanui et al. 2005). Copy number changes of *GDI1* correlate with the severity of X-linked mental retardation in males, with female carriers preferentially inactivating the mutant allele (Vandewalle et al. 2009). Somatic genomic deletions downstream from *IRS4* lead to its overexpression in lung squamous carcinoma (Weischenfeldt et al. 2017). Males with partial X disomy due to translocation of the distal long arm of the X Chromosome (Xq28) to the long arm of the Y

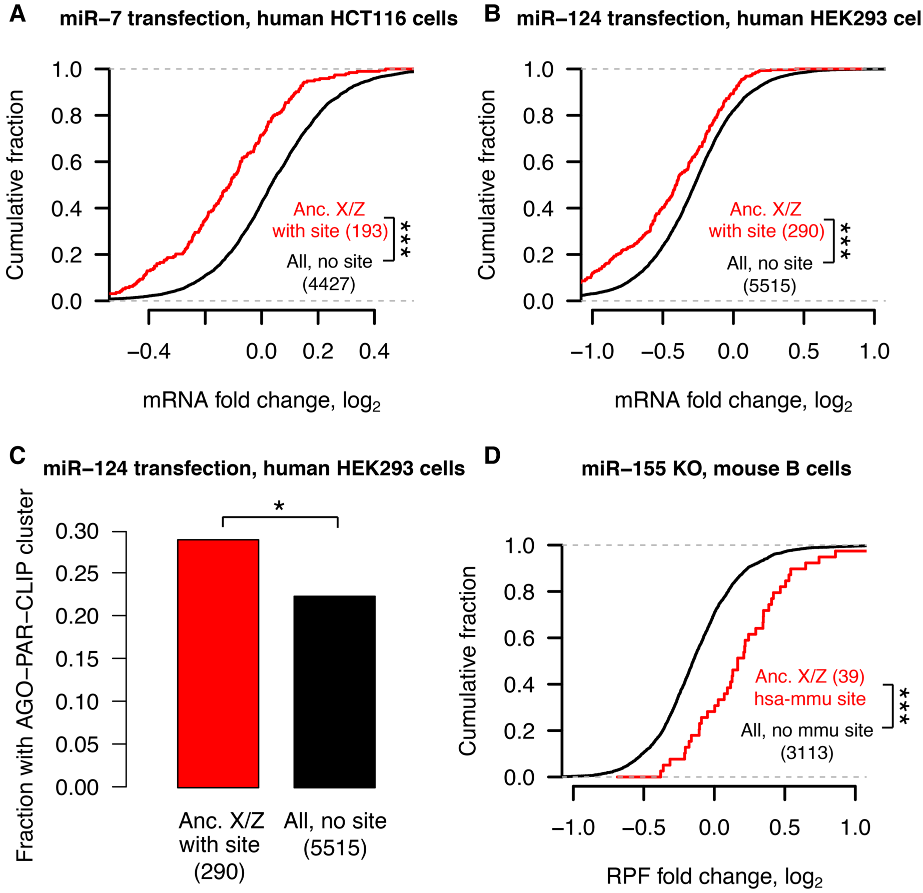


Figure 6. Analyses of experimental data sets validate miRNA target site function. Responses to transfection (*A–C*) or knockout (*D*) of indicated miRNAs in human (*A–C*) or mouse (*D*) cell types. Each panel depicts corresponding changes in mRNA levels (*A,B*), in fraction of Argonaute-bound genes (*C*), and in mRNA stability and translational efficiency as measured by ribosome protected fragments (RPF; *D*). In each case, X-linked genes and the human orthologs of Z-linked genes containing target sites with an assigned P_{CT} score (red) for the indicated miRNA were compared with all expressed genes lacking target sites (black); gene numbers are indicated in parentheses. (*A,B,D*) (***) $P < 0.001$, two-sided Kolmogorov–Smirnov test. (*C*) (*) $P < 0.05$, two-sided Fisher's exact test.

Chromosome show severe mental retardation and developmental defects (Lahn et al. 1994). Most genes in Xq28 are inactivated in 46,XX females but escape inactivation in such X;Y translocations, suggesting that increased dosage of Xq28 genes caused the cognitive and developmental defects. We anticipate that further studies will reveal additional examples of the deleterious effects of increases in gene dosage of X-Y pairs and X-inactivated genes.

We and others previously proposed that Y gene decay drove up-regulation of homologous X-linked genes in both males and females and that XCI subsequently evolved at genes sensitive to increased expression from two active X-linked copies in females (Ohno 1967; Jegalian and Page 1998). Our finding that X-inactivated genes have higher miRNA conservation scores than X escape genes is consistent with this aspect of the model. However, recent studies indicating heterogeneity in dosage sensitivity between classes of mammalian X- or avian Z-linked genes (Pessia et al. 2012; Bellott et al. 2014, 2017; Zimmer et al. 2016), combined with the present finding that these dosage sensitivities existed on the ancestral autosomes, challenge the previous assumption of a single evolutionary pathway for all sex-linked genes.

We therefore propose a revised model of X-Y and Z-W evolution in which the ancestral autosomes that gave rise to the mammalian and avian sex chromosomes contained three (or two, in the case of birds) classes of genes with differing dosage sensitivities (Fig. 7A,B). For ancestral genes with high dosage sensitivity, Y or W gene decay would have been highly deleterious, and thus the Y- or W-linked genes were retained. According to our model, these genes' high dosage sensitivity also precluded up-regulation of the X- or Z-linked homolog and, in mammals, subsequent X-inactivation; indeed, their X-linked homologs continue to escape XCI (Bellott et al. 2014). For ancestral mammalian genes of intermediate dosage sensitivity, Y gene decay did occur and was accompanied or followed by compensatory up-regulation of the X-linked homolog in both sexes; the resultant increased expression in females was deleterious and led to the acquisition of XCI. Ancestral mammalian genes of low dosage sensitivity continued to escape XCI following Y decay; heterogeneity in X up-regulation may further subdivide such genes (Fig. 6A). These genes' dosage insensitivity set them apart biologically, and evolutionarily, from the other class of X-linked genes escaping XCI—those with a surviving Y homolog.

Our revised model relates preexisting, gene-by-gene heterogeneities in dosage sensitivity to the outcomes of sex chromosome evolution. However, the suppression of X-Y recombination did not occur on a gene-by-gene basis, instead initiating Y gene decay and subsequent dosage compensation through a series of large-

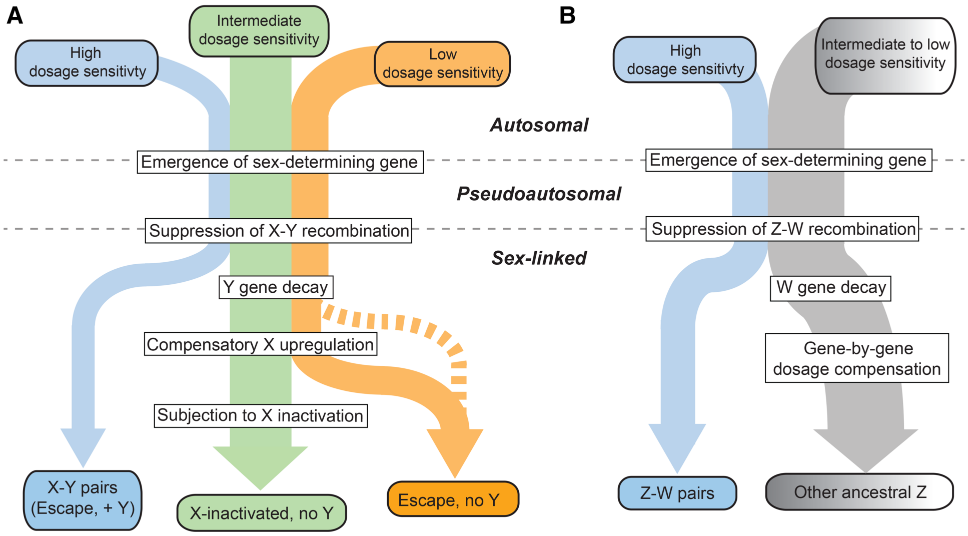

A
High dosage sensitivty
Intermediate dosage sensitivity
Low dosage sensitivity
Autosomal
Emergence of sex-determining gene
Pseudoautosomal
Suppression of X-Y recombination
Sex-linked
Y gene decay
Compensatory X upregulation
Subjection to X inactivation
X-Y pairs (Escape, + Y)
X-inactivated, no Y
Escape, no Y
B
High dosage sensitivty
Intermediate to low dosage sensitivity
Emergence of sex-determining gene
Suppression of Z-W recombination
W gene decay
Gene-by-gene dosage compensation
Z-W pairs
Other ancestral Z

Figure 7. An evidence-based model of preexisting heterogeneities in dosage sensitivity shaping mammalian and avian sex chromosome evolution. In this model, preexisting heterogeneities in dosage sensitivity determined the trajectory of Y/W gene loss in both mammals and birds, and of subsequent X-inactivation in mammals and dosage compensation in birds. Colored arrow widths are scaled approximately to the number of ancestral genes in each class. (*A*) The dashed orange line represents the possibility that a subset of X-linked genes may have not undergone compensatory X up-regulation following Y gene decay. (*B*) Ancestral Z genes with no W homolog follow a gradient of preexisting dosage sensitivity (*top*; gray to white), which determined the degree of dosage compensation following W gene loss (*bottom*; gray to white).

scale inversions encompassing many genes (Lahn and Page 1999). The timings and boundaries of these evolutionary strata varied among mammalian lineages, thus leading to unique chromosome-scale evolutionary dynamics across mammals. These large-scale changes would have then allowed for genic selection to take place according to the preexisting dosage sensitivities outlined above. In this way, the course of sex chromosome evolution in mammals is a composite of (1) preexisting, gene-by-gene dosage sensitivities and (2) the manner in which the history of the X and Y unfolded in particular lineages via discrete, large-scale inversions.

In this study, we have focused on classes of ancestral X-linked genes delineated by the survival of a human Y homolog or by the acquisition of XCI in humans, but such evolutionary states can differ among mammalian lineages and species. In mouse, for instance, both Y gene decay (Bellott et al. 2014) and the acquisition of X-inactivation (Yang et al. 2010) are more complete than in humans or other mammals, as exemplified by *RPS4X*, which retains a Y homolog and continues to escape XCI in primates but has lost its Y homolog and is subject to XCI in rodents. These observations could be explained by shortened generation times in the rodent lineage, resulting in longer evolutionary times, during which the forces leading to Y gene decay and the acquisition of X-inactivation could act (Ohno 1967; Charlesworth and Crow 1978; Jegalian and Page 1998). Another case of lineage differences involves *HUWE1*, which lacks a Y homolog and is subject to XCI in both human and mouse but retains a functional Y homolog in marsupials, where it continues to escape XCI. In the future, more complete catalogs of X-inactivation and escape in additional mammalian lineages would make it possible to examine whether analogous, preexisting dosage sensitivities differentiate the three classes of X-linked genes (X-Y pairs, X-inactivated genes, and X escape genes) in other species.

Previous studies have sought evidence of X-linked up-regulation during mammalian sex chromosome evolution using comparisons of gene expression levels between the whole of the X Chromosome and all of the autosomes, with equal numbers of studies rejecting or finding evidence consistent with up-regulation (Xiong et al. 2010; Deng et al. 2011; Kharchenko et al. 2011; Julien et al. 2012; Lin et al. 2012). This is likely due to gene-by-gene heterogeneity in dosage sensitivities that resulted in a stronger signature of up-regulation at more dosage-sensitive genes (Pessia et al. 2012). Similarly, studies of Z-linked gene expression in birds provide evidence for the gene-by-gene nature of Z dosage compensation, as measured by comparisons of gene expression levels between ZZ males and ZW females (Itoh et al. 2007; Mank and Ellegren 2009; Uebbing et al. 2015), and indicate a stronger signature of dosage compensation at predicted dosage-sensitive genes (Zimmer et al. 2016). By showing that such dosage sensitivities existed on the ancestral autosomes and consist of sensitivity to both increases and decreases, our findings highlight an additional aspect of dosage compensation that affects both birds and mammals.

In addition to revealing similarities between mammals and birds, our study provides a view of dosage compensation that highlights post-transcriptional regulatory mechanisms, pointing to specific noncoding sequences with known mechanisms (miRNA target sites) functioning across evolutionary time. A recent study in birds showed a role for a Z-linked miRNA, miR-2954-3p, in dosage compensation of some Z-linked genes (Warnefors et al. 2017). Our study suggests an additional, broader role for miRNA targeting, with hundreds of different miRNAs acting to tune gene dosage both before and during sex chromosome evolution. Furthermore, our finding of greater conserved miRNA targeting of X-inactivated genes relative to X escape genes shows that it is possible to predict the acquisition of a transcriptional regulatory state (XCI) during sex chromosome evolution on the basis of a preexisting, post-transcriptional regulatory state. Perhaps additional post-transcriptional regulatory mechanisms and their associated regulatory elements will be shown to play roles in mammalian and avian dosage compensation.

Recent work has revealed that the sex-specific chromosome—the Y in mammals and the W in birds—convergently retained dosage-sensitive genes with important regulatory functions (Bellott et al. 2014, 2017). Our study, by reconstructing the ancestral state of post-transcriptional regulation, provides direct evidence that such heterogeneity in dosage sensitivity existed on the ancestral autosomes that gave rise to the mammalian and avian sex chromosomes. This heterogeneity influenced both survival on the sex-specific chromosomes in mammals and birds and the evolution of XCI in mammals. Thus, two independent experiments of nature offer empirical evidence that modern-day amniote sex chromosomes were shaped, during evolution, by the properties of the ancestral autosomes from which they derive.

Methods

Statistics

Details of all statistical tests (type of test, test statistic, and *P*-value) used in this article are provided in Supplemental Table S1.

Human genic copy number variation

To annotate gene deletions and duplications, we used data from the ExAC (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/cnv/), which consists of autosomal genic duplications and deletions (both full and partial) called in 59,898 exomes (Ruderfer et al. 2016). Further details are provided in the Supplemental Methods in the section titled "Human genic copy number variation." These gene assignments are provided in Supplemental Table S2.

X- and Z-linked gene sets

We utilized our previous reconstructions of the ancestral mammalian X (Bellott et al. 2014) and avian Z (Bellott et al. 2017) Chromosomes, as well as information on multicopy and ampliconic X-linked genes (Mueller et al. 2013) and XCI status in humans (Balaton et al. 2015) to delineate classes of X- and Z-linked genes. Further details are provided in Supplemental Methods under the sections titled "X-linked gene sets" and "Z-linked gene sets." Information on X-linked genes is provided in Supplemental Table S3. Information on Z-linked genes is provided in Supplemental Table S4.

miRNA target sites

Precalculated P_{CT} scores for all gene–miRNA family interactions (http://www.targetscan.org/vert_71/vert_71_data_download/Summary_Counts.all_predictions.txt.zip) and site-wise alignment information (http://www.targetscan.org/vert_71/vert_71_data_download/Conserved_Family_Info.txt.zip) were obtained from TargetScan Human v7. Details on the filtering of miRNAs and resampling-based assessment of P_{CT} scores are provided in Supplemental Methods in the section titled "microRNA target site P_{CT} scores." Details regarding analysis of human–chicken or human–*Anolis* conserved sites, as well as approaches to control for background conservation, are provided in Supplemental Methods in the section titled "Human-chicken conserved microRNA target sites."

Variation in within-UTR conservation bias

To address the possibility that nonuniformity in regional 3′ UTR conservation could artificially inflate or deflate conservation scores of certain target sites, we implemented a step-detection algorithm to segment 3′ UTRs into regions of homogeneous background conservation and calculated miRNA site conservation relative to these smaller regions. These regionally normalized scores, corresponding to all gene–miRNA interactions, are provided in Supplemental Table S5. Details of the step-detection algorithm are provided in Supplemental Methods in the section titled "Variation in within-UTR conservation bias."

Logistic regression

Logistic regression models were constructed using the function "multinom" in the R package "nnet" (Venables and Ripley 2002). We used previously published values for known factors in the survival of Y-linked (Bellott et al. 2014) and W-linked (Bellott et al. 2017) genes except for human expression breadth, which we recalculated using data from the GTEx Consortium v6 data release (The GTEx Consortium 2015). Briefly, kallisto was used to estimate transcript per million (TPM) values in the 10 male samples with the highest RNA integrity numbers (RINs) from each of 37 tissues, and expression breadth across tissues was calculated as described by Bellott et al. (2014), using median TPM values for each tissue.

Assessing Z-linked dosage compensation using cross-species RNA-sequencing data

Raw data from Marin et al. (2017) was obtained, and kallisto and limma/voom were used for abundance quantification and differential expression, respectively. Further details are provided in the Supplemental Methods in the section titled "Assessing Z-linked dosage compensation using cross-species RNA-sequencing data."

Gene expression profiling and crosslinking data sets

Fold-changes in mRNA expression and targets of Argonaute as determined by high-throughput CLIP were obtained from a variety of publicly available data sets. Further details are provided in Supplemental Methods in the section titled "Gene expression profiling and crosslinking datasets." All fold-changes and CLIP targets are provided in Supplemental Table S6.

Software availability

A custom Python (RRID:SCR_008394) script utilizing Biopython (RRID:SCR_007173) was used to generate shuffled miRNA family seed sequences. Identification of miRNA target site matches using shuffled seed sequences was performed using the "targetscan_70.pl" Perl script (http://www.targetscan.org/vert_71/vert_71_data_download/targetscan_70.zip). 3′ UTR segmentation was performed with the "plot_transitions.py" Python script. Code is available at https://github.com/snaqvi1990/Naqvi17-code and as Supplemental Code.

Acknowledgments

We thank V. Agarwal, S. Eichorn, S. McGeary, and D. Bartel for assistance with the TargetScan database and helpful discussions; A. Godfrey for updated human–chicken orthology information; and A. Godfrey, J. Hughes, and H. Skaletsky for critical reading of the manuscript. This work was supported by the National Institutes of Health and the Howard Hughes Medical Institute. S.N. was supported under a research grant by Biogen.

Author contributions: S.N., D.W.B., and D.C.P. designed the study. S.N. performed analyses with assistance from D.W.B. K.S.L. developed and implemented the step-detection algorithm. S.N. and D.C.P. wrote the paper.

References

Balaton BP, Cotton AM, Brown CJ. 2015. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. *Biol Sex Differ* **6:** 35.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136:** 215–233.

Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, Brown LG, Rozen S, Warren WC, Wilson RK, et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466:** 612–616.

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, Koutseva N, Zaghlul S, Graves T, Rock S, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508:** 494–499.

Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T, et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat Genet* **49:** 387–394.

Berletch JB, Ma W, Yang F, Shendure J, Noble WS, Disteche CM, Deng X. 2015. Escape from X inactivation varies in mouse tissues. *PLoS Genet* **11:** e1005079.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434:** 400–404.

Charlesworth B, Crow JF. 1978. Model for evolution of Y chromosomes and dosage compensation. *Proc Natl Acad Sci* **75:** 5618–5622.

Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, Hillier LW, Schlesinger F, Davis CA, Reinke VJ, Gingeras TR, et al. 2011. Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat Genet* **43:** 1179–1185.

Friedman RC, Farh KK-H, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19:** 92–105.

The GTEx Consortium. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348:** 648–660.

Hemara-Wahanui A, Berjukow S, Hope CI, Dearden PK, Wu S-B, Wilson-Wheeler J, Sharp DM, Lundon-Treweek P, Clover GM, Hoda J-C, et al. 2005. A *CACNA1F* mutation identified in an X-linked retinal disorder shifts the voltage dependence of $Ca_v1.4$ channel activation. *Proc Natl Acad Sci* **102:** 7553–7558.

Huang N, Lee I, Marcotte EM, Hurles ME. 2010. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6:** e1001154.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483:** 82–86.

Itoh Y, Melamed E, Yang X, Kampf K, Wang S, Yehya N, Van Nas A, Replogle K, Band MR, Clayton DF, et al. 2007. Dosage compensation is less effective in birds than in mammals. *J Biol* **6:** 2.

Jegalian K, Page DC. 1998. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394:** 776–780.

Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schütz F, Daish T, Grützner F, Kaessmann H. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol* **10:** e1001328.

Kaiser VB, Zhou Q, Bachtrog D. 2011. Nonrandom gene loss from the *Drosophila miranda* neo-Y chromosome. *Genome Biol Evol* **3:** 1329–1337.

Kharchenko PV, Xi R, Park PJ. 2011. Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nat Genet* **43:** 1167–1169.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286:** 964–967.

Lahn BT, Ma N, Breg RW, Stratton R, Surti U, Page DC. 1994. Xq-Yq interchange resulting in supernormal X-linked gene expression in severely retarded males with 46,XYq- karyotype. *Nat Genet* **8:** 243–250.

Lin F, Xing K, Zhang J, He X. 2012. Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. *Proc Natl Acad Sci* **109:** 11752–11757.

Lindgren AM, Hoyos T, Talkowski ME, Hanscom C, Blumenthal I, Chiang C, Ernst C, Pereira S, Ordulu Z, Clericuzio C, et al. 2013. Haploinsufficiency of *KDM6A* is associated with severe psychomotor retardation, global growth restriction, seizures and cleft palate. *Hum Genet* **132:** 537–552.

Mank JE, Ellegren H. 2009. All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity* **102:** 312–320.

Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Brüning T, Mössinger K, et al. 2017. Convergent origination of a *Drosophila*-like dosage compensation mechanism in a reptile lineage. *Genome Res* **27:** 1974–1987.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat Genet* **45:** 1083–1087.

Nanda I, Shan Z, Schartl M, Burt DW, Koehler M, Nothwang H, Grützner F, Paton IR, Windsor D, Dunn I, et al. 1999. 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat Genet* **21:** 258–259.

Ohno S. 1967. *Sex chromosomes and sex-linked genes*. Springer-Verlag, Berlin.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194–197.

Pessia E, Makino T, Bailly-Bechet M, McLysaght A, Marais GA. 2012. Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc Natl Acad Sci* **109:** 5346–5351.

Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, Seitz H. 2017. The number of biologically relevant microRNA targets has been largely over-estimated. *Genome Res* **27:** 234–245.

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M. 2005. The DNA sequence of the human X chromosome. *Nature* **434:** 325–337.

Ruderfer DM, Hamamsy T, Lek M, Karczewski KJ, Kavanagh D, Samocha KE, Exome Aggregation Consortium, Daly MJ, MacArthur DG, Fromer M, et al. 2016. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* **48:** 1107–1111.

Simon D, Laloo B, Barillot M, Barnetche T, Blanchard C, Rooryck C, Marche M, Burgelin I, Coupry I, Chassaing N, et al. 2010. A mutation in the 3′-UTR of the *HDAC6* gene abolishing the post-transcriptional regulation mediated by hsa-miR-433 is linked to a new form of dominant X-linked chondrodysplasia. *Hum Mol Genet* **19:** 2015–2027.

Skaletsky H, Kuroda-kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825–838.

Tukiainen T, Villani A-C, Yen A, Rivas MA, Marshall JL, Satija R, Aguirre M, Gauthier L, Fleharty M, Kirby A, et al. 2017. Landscape of X chromosome inactivation across human tissues. *Nature* **550:** 244–248.

Uebbing S, Konzer A, Xu L, Backström N, Brunström B, Bergquist J, Ellegren H. 2015. Quantitative mass spectrometry reveals partial translational regulation for dosage compensation in chicken. *Mol Biol Evol* **32:** 2716–2725.

Vandewalle J, Van Esch H, Govaerts K, Verbeeck J, Zweier C, Madrigal I, Mila M, Pijkels E, Fernandez I, Kohlhase J, et al. 2009. Dosage-dependent severity of the phenotype in patients with mental retardation due to a recurrent copy-number gain at Xq28 mediated by an unusual recombination. *Am J Hum Genet* **85:** 809–822.

Venables WN, Ripley BD. 2002. *Modern applied statistics with S*, 4th ed. Springer, New York.

Warnefors M, Mossinger K, Halbert J, Studer T, VandeBerg JL, Lindgren I, Fallahshahroudi A, Jensen P, Kaessmann H. 2017. Sex-biased microRNA expression in mammals and birds reveals underlying regulatory mechanisms and a role in dosage compensation. *Genome Res* **27:** 1961–1973.

Watson JM, Spencer JA, Riggs AD, Graves JA. 1990. The X chromosome of monotremes shares a highly conserved region with the eutherian and marsupial X chromosomes despite the absence of X chromosome inactivation. *Proc Natl Acad Sci* **87:** 7125–7129.

Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stütz AM, Waszak SM, Bosco G, Halvorsen AR, Raeder B, et al. 2017. Pan-cancer analysis of somatic copy-number alterations implicates *IRS4* and *IGF2* in enhancer hijacking. *Nat Genet* **49:** 65–74.

White MA, Kitano J, Peichel CL. 2015. Purifying selection maintains dosage-sensitive genes during degeneration of the three-spine stickleback Y chromosome. *Mol Biol Evol* **32:** 1981–1995.

Xiong Y, Chen X, Chen Z, Wang X, Shi S, Wang X, Zhang J, He X. 2010. RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat Genet* **42:** 1043–1047.

Yang F, Babak T, Shendure J, Disteche CM. 2010. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20:** 614–622.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res* **44:** D710–D716.

Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, Gilbert MT, Zhang G. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346:** 1246338.

Zimmer F, Harrison PW, Dessimoz C, Mank JE. 2016. Compensation of dosage-sensitive genes on the chicken Z chromosome. *Genome Biol Evol* **8:** 1233–1242.

Received September 21, 2017; accepted in revised form February 6, 2018.

Appendix 2. 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.

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.

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 more rapid for short-tailed mRNAs that had 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.

Highlights:

- mRNAs enter the cytoplasm with diverse intra- and intergenic 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

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 (Muhlemann 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 mRNA analysis 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 sequencing methods, each with distinct advantages, were initially developed to profile poly(A)-tail lengths. One of these is PAL-seq (poly(A)-tail-length profiling by sequencing), which 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-adenosine residues (Chang et al., 2014; Lim et al., 2016). Here, we developed PAL-seq 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 the 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. 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 a biotin-containing 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 provided measurements of abundance for mRNAs of each time interval. Spike-ins of RNA standards 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 each performed using each of two independently passaged 3T3 cell lines. Unless stated otherwise, all figures show the results obtained for cell line 1. Nonetheless, the results of the two cell lines were highly 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 (Figure S2A–D; $R_s = 0.83$–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), mRNAs collected after the shortest labeling period (40 min) had the longest poly(A)-tail lengths, with a 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 towards the steady-state distribution (median length of 91 nt), with results from the 8 h period most closely resembling those of the steady state 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 (Figure 1C–D).

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 tail lengths 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 were 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 27–36 min, depending on the experiment, a range 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 (Figure S3B–D, median 2.1 h for mRNAs of the 3T3 cell line 2, 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 (made possible by adding standards to each sample prior to library preparation).

Previous 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 has been observed (Subtelny et al., 2014), despite the established role of poly(A) tails in conferring mRNA stability. Our current datasets, which provided the opportunity make this comparison using half-life and tail-length measurements determined in the same study 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 ribosome-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 RPG mRNAs and other 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 IEG mRNAs and other 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, suggesting that their degradation and deadenylation rates were incongruous. *Rassf1*, *Mat2a*, *Serpine1*, and two *Gadd45* paralogs are known or suspected substrates for either nonsense-mediated decay (NMD) or other pathways that recruit UPF1 (Forrest et al., 2004; Tani et al., 2012; Park and Maquat, 2013; Bresson et al., 2015; Nelson et al., 2016). 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 (Muhlemann 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).

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 (Figure 1C, Figure 3, and Figure S3F), with the spread of the 5th to 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, respectively, values ± s.d.), 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 to 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, respectively, values ± s.d.), 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 in the 40 min sample. These c.v. values were reproducible between biological replicates and had little correspondence with mRNA half-life (Figure S3I–J). 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 abundance 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). In contrast to mRNA half-lives, which are fit directly to the approach to steady state, other rates are not as simple to calculate. 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, 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 is no longer protected from decapping, and 5) the decapping rate of short-tailed mRNAs (with decapping assumed to trigger rapid decay of the mRNA body). Therefore, we developed a mathematical model to determine, for mRNAs of 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 and Table S2), an approach resembling that used to model 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. Decapping, with rate constant k_2, can occur alongside deadenylation and monotonically increases as the poly(A)-tails get shorter. Because the body of each decapped mRNA is rapidly degraded (instantly in our model), decapping reduces the abundance of mRNAs in the pool. Note that although this model parameter is named “decapping” based on the idea that mRNA decay proceeds primarily through decapping and subsequent 5′-to-3′ decay, decay of a short-tailed mRNA body by other mechanisms would also be consistent with our model.

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 and Table S2). The decapping rate constant followed a logistic function, which accelerated decapping 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 (Figure S4D–E). These results indicated that, for endogenous mRNAs in 3T3 cells, either a single deadenylase complex dominates—as recently proposed for mRNAs with tail lengths ≤ 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 decapping 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 and 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 decapping were reproducible between biological replicates and robust to parameter initialization as well as multinomial sampling (bootstrap analysis) (Figure S4G–J).

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 metabolically characterized 2778 mammalian 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 0.03 nt/min (Figure 5A). Concordant with our direct analysis of the primary data, which revealed a strong correspondence between mRNA half-life and pre-steady-state mean tail length, thereby implying 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 decapping rates for all measured genes at all tail lengths and thereby infer the tail lengths at which mRNAs were decapped and degraded (Figure 5B). This analysis indicated that nearly all decapping occurred after the tail lengths fell below 100 nt, which agreed with previous analyses of reporter genes (Yamashita et al., 2005). Decapping greatly accelerated as tail lengths fell below 50 nt (with > 92% of mRNAs decapped below this length), a length less than the 54 nt footprint of two cytoplasmic poly(A)-binding protein (PABPC) molecules (Baer and Kornberg, 1983; Yi et al., 2018). However, most mRNA molecules (> 55%) were not decapped 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 the mean tail length of decapping for mRNAs of each gene, the results generally concurred with those observed for all mRNA combined, with mRNAs from most genes decapped at short mean tail lengths (Figure 5C, > 97% decapped at mean tail length < 50 nt and > 69% decapped 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 *H2afx*, *Mat2a*, *Gadd45b*, and *Marveld1*, degrading at a mean tail length of 75, 70, 62, and 59 nt, respectively. The estimates of mean decapping tail lengths together with initial tail lengths and deadenylation rate constants enabled estimates of the time required to reach the mean tail length of decapping, which corresponded to lifetime slightly better than did the deadenylation rate constants on their own to half-life (Figure S5A–B, R_s = –0.96 and –0.95, respectively.)

Once tails reached a short length, the decapping rate constants of short-tailed mRNAs varied widely, with short-tailed mRNAs from some genes undergoing decapping 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). If general for other mRNAs, a more rapid degradation of short-tailed mRNAs that had been more rapidly deadenylated would help prevent buildup of short-tailed isoforms of rapidly 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). As reported for *Fos*, 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 this more rapid clearing of mRNAs that had been more rapidly deadenylated, we investigated whether this phenomenon was able to prevent a large build-up 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 (Figure 6A and Figure S5C, $R_s = -0.56$). Nonetheless, the buildup of short-tailed isoforms of these unstable RNAs usually failed to exceeded 30% of all transcripts (Figure 6A).

This preferential buildup of short-tailed isoforms of unstable RNAs was 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 0–2 nt, whereas long-lived mRNAs (half-lives > 10 h) were depleted of 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 (Figure 6C, D and Figure 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 suggesting that uridylation preferentially occurs 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 Figure 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 (Figure S5F–H), 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) (Figure 6B and Figure S5I–J), perhaps due to an initial lag in assembling deadenylation machinery as mRNAs enter the cytoplasm, which would the 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 5B) 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 for short- and long-lived mRNAs, resepectively), 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 decayed. 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 Populations of Synchronous mRNAs

Our continuous-labeling experiments were designed to measure dynamics of mRNA metabolism in an unperturbed cellular environment. However, this framework required deadenylation and decapping 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 the end behavior of mRNAs 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 the 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, the tail lengths of labeled mRNAs progressively decreased after transcriptional inhibition, with median tail 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 the mean tail-length distributions peaked between 45–50 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). This 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 the 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 behavior. 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 1h 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 ($R_s = 0.88$) provided compelling support for our conclusion that the vast majority of mRNAs are primarily degraded through deadenylation-linked mechanisms.

To further analyze 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. The minimal degradation of long-tailed molecules disfavored the idea that preferential degradation of long-tailed mRNAs might help explain the shift in tail lengths observed with increasing time intervals in the continuous-labeling experiment. 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, decapping, 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 2687 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 results and conclusions from both experiments as well as from our analytical framework.

Discussion

Previous mechanistic studies of mRNA turnover 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 examination of the dynamics of deadenylation and degradation for mRNAs from thousands of endogenous genes provided a more comprehensive resource for deriving the principles of cytoplasmic mRNA metabolism. Initial analyses of our data 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 decapping and destruction of the mRNA body. With respect to the reason that short tail lengths trigger decay, our estimates of the tail lengths at which most mRNAs decay supported the prevailing view that loss of PABPC binding to the poly(A) tail enhances decay, with the idea that occupancy is somewhat lower on tails too short for cooperative binding of a PABPC dimer (tail lengths ~50 nt) and much lower on those too short for efficient binding of a single PABPC molecule (tail lengths ~20 nt).

This 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 decapping rates functionally integrates mRNA turnover into a single process to ensure that mRNAs that are rapidly deadenylated are also rapidly cleared from the cell, which enables the large range in deadenylation rate constants to impart an equally large range in mRNA stabilities. 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. 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 decapping rates.

The large differences observed for both deadenylation and decapping 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 the presence or absence of these 3′-UTR sites helps to specify the differences (Mauxion et al., 2009; Muhlemann 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, despite the prevalence of regulatory factors that bind to 3′-UTR sites, 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, our global analyses indicated that recently exported, long-tailed mRNAs typically undergo 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. In most analyses, however, we used our mathematical model of the kinetics of deadenylation and decay to capture critical features of mRNA metabolism missed by naïve exponential decay.

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 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 nt and 7–15 nt (Figure 6B). A model with uniform deadenylaiton 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 build-up 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. Several more parameters would be required to model a buildup of 7–15 nt tails, which 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. Modeling the extensive terminal uridylation of the short-tailed isoforms of short-lived mRNAs that accumulate (Figure 6C) might provide insight into the function of this modification. Previous studies report that terminal uridylation of the poly(A) tail stimulates decapping of mRNAs in *Aspergillus* and fission yeast, and of histone mRNAs in HeLa cells, and high-throughput studies link terminal uridylation more generally to mammalian mRNA stability (Mullen and

Marzluff, 2008; Rissland and Norbury, 2009; Morozov et al., 2010; Chang et al., 2014; Lim et al., 2014). Based on these previous findings and the results of our analyses, we speculate that uridylation of non-histone mammalian mRNAs is preferentially deployed to short-tailed isoforms of more rapidly deadenylated transcripts to promote more rapid decapping, which helps prevent a larger buildup of these short-tailed isoforms. Regardless of the role for uridylation, the observation of these buildups of preferentially uridylated 0–1 and 7–15 nt tails for short-lived mRNAs helps explain the observations that, compared to long-lived mRNAs, short-lived RNAs have both a higher fraction of uridylated tails and a higher fraction of short-tailed isoforms that are uridylated (Chang et al., 2014; Lim et al., 2014).

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 in which a G or occasionally another non-A nucleotide has been incorporated (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 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.

Acknowledgements

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.). 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.

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.

Bresson, S.M., Hunter, O.V., Hunter, A.C., and Conrad, N.K. (2015). Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. PLoS Genet *11*, e1005610.

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.

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.

Dölken, L., Ruzsics, Z., Radle, B., Friedel, C.C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P.*, et al.* (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., Villen, J., and Bartel, D.P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. Mol Cell *56*, 104-115.

Eisen, T.J., Eichhorn, S.W., Subtelny, A.O., and Bartel, D.P. (2019). MicroRNAs Cause Accelerated Decay of Short-Tailed Target mRNAs. bioRxiv.

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

Forrest, S.T., Barringhaus, K.G., Perlegas, D., Hammarskjold, M.L., and McNamara, C.A. (2004). Intron retention generates a novel Id3 isoform that inhibits vascular lesion formation. J Biol Chem *279*, 32897-32903.

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.

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 U S A *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.

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., 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.

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.

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., Seraphin, 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.

Muhlemann, 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.

Mullen, T.E., and Marzluff, W.F. (2008). Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5' to 3' and 3' to 5'. Genes Dev *22*, 50-65.

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*.

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.

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.

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.

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 e129.

Rissland, O.S., Mikulasova, A., and Norbury, C.J. (2007). Efficient RNA polyuridylation by noncanonical poly(A) polymerases. Mol Cell Biol *27*, 3612-3624.

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.

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.

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., Moldon, 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 e1088.

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*.

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 e1085.

Methods

Cell culture

Clonal 3T3 cell lines engineered to express miR-155 (cell line 1) or miR-1 (cell line 2) upon doxycycline treatment were previously described (Eichhorn et al., 2014). Cells were grown at 37°C in 5% CO_2 in DMEM supplemented with 10% BCS (Sigma-Aldrich) and 2 μg/mL puromycin. For metabolic-labeling time courses, cells from each line were plated onto 500 cm^2 plates at 6.6 million cells per plate and cultured for two days such that they reached ~70–80% confluency, at which point growth media was supplemented with 5-ethynyluridine (5EU, Jena Biosciences) (Jao and Salic, 2008) at a final concentration of 400 μM. After the desired labeling intervals cells were harvested (Figure 2A). Four plates were harvested for each 40 min time interval, three plates for each 1 h time interval, and two plates for each other time interval. A plate that had never received 5EU was harvested in parallel for each condition.

Cells were harvested at 4°C, washed twice with 50 mL ice-cold 9.5 mM PBS, pH 7.3 containing 100 μg/mL cycloheximide and then used to prepare cytoplasmically enriched lysate as described (Subtelny et al., 2014). An aliquot of cleared lysate was flash frozen for use in ribosome profiling (Eisen et al., 2019), and the rest of the lysate was added to 5 volumes of TRI reagent (Ambion) and frozen at –80°C. Samples stored in TRI reagent were thawed at room temperature, and RNA was purified according to the manufacturer's protocol and used for RNA-seq or PAL-seq v2.

RNA standards

Two sets of tail-length standards (set 1 and set 3, Table S3) were described previously (standard mix 2 and standard mix 1) (Subtelny et al., 2014). The other set of standards (set 2, Table S3) was prepared based on a 705 nt fragment of the *Renilla* luciferase mRNA, which was transcribed and gel purified as described (Subtelny et al., 2014) and then capped using a Vaccinia capping system (2000 µL reaction containing 500 µg RNA, 1000 U Vaccinia capping enzyme (NEB), 1X Capping Buffer (NEB), 0.1 mM S-adenosyl methionine, 0.5 mM GTP, 50 nM [α–^{32}P]-GTP, 2000 U SUPERaseIn (ThermoFisher) at 37°C for 1 h), monitoring the amount of incorporated radioactivity to ensure that capping was quantitative. Following the capping reaction, the 2′,3′ cyclic phosphate at the 3′ end was removed using T4 polynucleotide kinase (Subtelny et al., 2014). The capped, dephosphorylated product was joined by splinted ligation to each of seven different poly(A)-tailed barcode oligonucleotides (Subtelny et al., 2014). These seven 3′ ligation partners included 110 and 210 nt poly(A) oligonucleotides prepared as described (Subtelny et al., 2014), and five gel-purified synthetic oligonucleotides (IDT), one with a 10 nt poly(A) tract and the other four with a 29 nt poly(A) tract followed by either A, C, G, or U. Ligation products were gel purified, mixed in desired ratios, with final ratios of the different-sized species confirmed by analysis on a denaturing polyacrylamide gel.

Short and long standards were used to monitor enrichment of 5EU-containing fragmented RNA or non-fragmented RNA, respectively. Short 5EU standards were prepared by in vitro transcription of annealed DNA oligos to produce a 30 nt and 40 nt RNA, with the latter containing a single 5EU (Table S3). In vitro transcription was performed with the MEGAscript T7 transcription kit (ThermoFisher) according to the manufacturer’s protocol, except UTP was replaced with 5-ethynyluridine-triphosphate

(Jena Biosciences) when transcribing the 40 nt RNA. Long standards were prepared by in vitro transcription of sequences encoding firefly luciferase and GFP using the MEGAscript T7 transcription kit and 0.1 µM PCR product as the template. When transcribing *GFP*, a 20:1 ratio of UTP to 5-ethynyluridine-triphosphate was used. Short and long standards were gel purified and stored at –80°C. Prior to use, a portion of each standard was cap-labeled and gel purified again, which enabled measurement of the recovery of the 5EU-containing standard relative to that of the uridine-only standard.

Three 28–30 nt RNAs (Table S3) were synthesized (IDT) for use as quantification standards in RNA-seq. These standards were gel purified, and 0.1 fmol of each was added to each sample immediately prior to library preparation.

Biotinylation of 5EU labeled RNA

The RNA-seq libraries analyzed in this study were from fragmented RNAs, size selected to match ribosome-profiling libraries (Eisen et al., 2019). For these libraries, poly(A) RNA was purified from 50 µg total RNA of the 40 min, 1, 2, and 4 h samples and 25 µg total RNA of the 8 h sample using oligo(dT) Dynabeads (ThermoFisher) according to manufacturer's protocol. RNA was fragmented and 27–33 nt fragments were isolated as described (Subtelny et al., 2014), short standards that monitored 5EU enrichment were added, and then Cu(II) catalysis was used to biotinylate 5EU in a 20 µL reaction containing 50 mM HEPES, pH 7.5, 4 mM disulfide biotin azide (Click Chemistry Tools), 2.5 mM $CuSO_4$, 2.5 mM Tris(3-hydroxypropyltriazolylmethyl)amine (THPTA, Sigma-Aldrich), and 10 mM sodium ascorbate, incubated at room temperature for 1 h. Reactions were stopped with 5 mM EDTA and then extracted with phenol–chloroform (pH 8.0).

For the steady-state samples, 5 μg of RNA from the 40 min sample was poly(A) selected and fragmented, and size-selected 27–33 nt fragments were carried forward without enriching for 5EU.

For PAL-seq v2, long standards used to monitor 5EU enrichment and recovery were added to total RNA (using a 1:10 ratio of 5EU-containing standard to non-5EU-containing standard), and samples were click labeled as above in reactions with 2.5 μg/μL RNA. For samples from the cell line 1 time course, click reactions were performed with 500, 500, 250, 200, or 100 μg total RNA for the 40 min, 1 h, 2 h, 4 h, or 8 h samples. For samples from the cell line 2 time course, click reactions were performed with 800, 525, 350, or 200 μg total RNA for the 40 min, 1 h, 2 h, or 4 h, respectively. For both cell lines, the steady-state samples did not undergo click reactions or pull-down.

Purification of biotinylated RNA

For RNA-seq, Dynabeads MyOne Streptavidin C1 beads (ThermoFisher) for each set of samples were combined and batch washed, starting with 200 μL of beads per reaction. Beads were washed twice with 1X B&W buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl and 0.005% Tween-20), twice with solution A (0.1 M NaOH, 50 mM NaCl), twice with solution B (0.1 M NaCl), and then twice with water, using for each wash a volume equal to that of the initial bead suspension. Following the last wash, beads were resuspended in an initial bead volume of 1X high salt wash buffer (HSWB, 10 mM Tris-HCl, pH 7.4, 1 mM EDTA, 0.1 M NaCl, 0.01% Tween-20) supplemented with 0.5 μg/mL yeast RNA (ThermoFisher) and incubated at room temperature for 30 min with end-over-end rotation, again using a volume equal to that of the initial bead suspension. Beads were then washed three times with 200 µL 1X HSWB per reaction and split for each reaction during the last wash. After the wash was removed, sample RNA resuspended in 200 µL 1X HSWB was added to blocked beads and incubated with end-over-end rotation at room temperature for 30 min. Beads were washed twice with 800 µL 50°C water, incubating at 50°C for 2 min for each wash, and then twice with 800 µL 10X HSWB. RNA was eluted from beads by incubating with 200 µL 0.5 M tris(2-carboxyethyl)phosphine (TCEP, Sigma-Aldrich) at 50°C for 20 min with end-over-end rotation. The initial eluate was collected, and beads were resuspended in 150 µL water and eluted again, combining the two eluates for each sample. RNA from the eluate was then ethanol precipitated using linear acrylamide as a carrier.

Purifications of non-fragmented RNA were performed as above, except bead volumes were adjusted based on estimates of the amount of labeled RNA in each sample. For the cell line 1 samples, 292, 431, 410, 598, and 500 µL of beads were used for the 40 min, 1 h, 2 h, 4 h, and 8 h samples, respectively. For the cell line 2 samples, 467, 452, 575, and 598 µL streptavidin beads were used for the 40 min, 1 h, 2 h, and 4 h samples, respectively.

Pilot experiments designed to optimize the 5EU biotinylation and purification confirmed that RNAs containing at least one 5EU could be purified efficiently, with over 80% of a model RNA substrate containing a single 5EU becoming biotinylated in a 1 h reaction (Eisen et al., 2019). This high reaction efficiency was important for the RNA-seq samples, as RNA fragments from these libraries, generated to match ribosome-profiling samples (Eisen et al., 2019), were only ~30 nt long and estimated to typically contain at most a single 5EU. Indeed, for each of the three protocols, which started with either full-length RNA (PAL-seq) or fragmented RNA (RNA-seq), metabolically labeled RNA was substantially enriched above background (Eisen et al., 2019).

PAL-seq v2

This method starts with the same mRNA workup as PAL-seq v1 (Subtelny et al., 2014), except the design of the 3′ adapter allows for ligation to tails ending with a uridine nucleotide, as implemented in an improved version of TAIL-seq (Lim et al., 2016). PAL-seq v2 also includes a primer-extension reaction that occurs on the Illumina flowcell, with the goal of extending the sequencing primer all of the way through the poly(A) tail, so that the first sequencing read identifies both the mRNA and its cleavage-and-polyadenylation site, as in PAL-seq v1 (Subtelny et al., 2014). The poly(A)-tail length is then measured by direct sequencing of the poly(A) tail, as in TAIL-seq (Chang et al., 2014) (Figure S1A).

We used RNA standards of defined tail lengths to monitor library preparation, sequencing, and the computational pipeline for improved versions of PAL-seq and our implementation of TAIL-seq. Depletion of long-tailed sequences was the most prevalent source of measurement error. For TAIL-seq, this depletion seemed highly dependent on the sequencing protocol, with the best results obtained on a HiSeq machine in high-output mode using the v3 reagent kit.

Steady-state RNA (25 µg of unselected RNA from the 40 min sample) or half of the RNA eluted from each 5EU-selected sample was used to prepare PAL-seq libraries. Tail-length standard mixes (1 ng of set 1 and 2 ng of set 2 for each 5EU-selected sample, and twice these amounts for the steady-state sample), and trace 5′-radioiolabeled marker

RNAs (Table S3) were added to each sample to assess tail-length measurements and ligation outcomes, respectively. Polyadenylated ends including those with a terminal uridine were ligated to a 3′-biotinylated adapter DNA oligonucleotide (1.8μM) in the presence of two splint DNA oligonucleotides (1.25μM and 0.25μM for the U and A-containing splint oligos, respectively, Table S3) using T4 Rnl2 (NEB) in an overnight reaction at 18°C. Following 3′-adapter ligation the RNA was extracted with phenol–chloroform (pH 8.0), precipitated, resuspended in 1X RNA T1 sequence buffer (ThemoFisher), heated to 50°C for 5 min and then put on ice. RNase T1 was then added to a final concentration of 0.006 U/μL, and the reaction was incubated at room temperature for 30 min, followed by phenol–chloroform extraction and RNA precipitation. Precipitated RNA was captured on streptavidin beads, 5′ phosphorylated, and ligated to a 5′ adapter as described (Subtelny et al., 2014) but using a modified 5′ adapter sequence (Table S3). Following reverse transcription using SuperScript III (Invitrogen) with a barcode-containing DNA primer, cDNA was purified as described (Subtelny et al., 2014), except a 160–810 nt size range was selected. Libraries were amplified by PCR for 8 cycles using Titanium Taq polymerase according to the manufacturer's protocol with a 1.5 min combined annealing/extension step at 57°C. PCR-amplified libraries were purified using AMPure beads (Agencourt, 40 μL beads per 50 μL PCR, two rounds of purification) according to the manufacturer's instructions.

The use of a splinted ligation of the 3′ adapter to the poly(A) tail had the advantage of specifically ligating to mRNAs without the need to deplete ribosomes or other abundant RNAs. However, this approach was not suitable for acquiring measurements for mRNAs with tails that were either very short (< 8 nt) or extended by more than one uridine, because such tails would ligate less efficiently (or not at all) when using a splinted ligation to the 3′ adapter. To account for these mRNAs with either very short or highly modified tails, we implemented a protocol that used single-stranded (ss) ligation and different mRNA enrichment steps to prepared libraries from steady-state RNA isolated from each of the two cell lines. For each sample, 5 µg of total RNA was depleted of rRNA using RiboZero Gold HMR (Illumina) and further depleted of the 5.8s rRNA by subtractive hybridization. Subtractive hybridization was performed by mixing 2x SSC buffer (3M sodium chloride, 300mM sodium citrate, pH 7.0), total RNA, and 4.8µM of each 5.8s subtractive-hybridization oligo (Table S3) in a 50 µL reaction, heating the reaction to 70°C for 5 min, then cooling it at 1°C/min to 37°C to anneal the oligos to the RNA. During this cooling, 250 µL of Dynabeads MyOne Streptavidin C1 beads per sample (ThermoFisher) were washed twice with 1X B&W buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl and 0.005% Tween-20), twice with solution A (0.1 M NaOH, 50 mM NaCl), twice with solution B (0.1 M NaCl), and then resuspended in 50 µL of 2X B&W buffer. After cooling, the entire 50µL RNA/oligo mixture was added to 50 µL of washed beads, then incubated at room temperature for 15 min with end-over-end rotation. The sample was then magnetized and the supernatant was withdrawn and precipitated by adding 284 µL of water, 4 µL of 5 mg/mL linear acrylamide, and 1 mL of ice-cold 96% ethanol. After resuspension, RNA was ligated to a 3′ adapter containing four random-sequence nucleotides and an adenylyl group at its 5′ end (Table S3) in a 70 µl reaction containing 10 µM adapter, 1X T4 RNA Ligase Reaction Buffer (NEB), 20 U/µL T4 RNA Ligase 2 truncated KQ (NEB), 0.3 U/µL SUPERaseIn (ThermoFisher), and 20% PEG 8000. The reaction was incubated at 22°C overnight and then stopped by addition of EDTA (3.5 mM final after bringing the reaction to 400 µL with water). RNA was phenol–chloroform extracted, precipitated, and subsequent library preparation was as for the splinted-ligation libraries.

PAL-seq v2 libraries were sequenced on an Illumina HiSeq 2500 operating in rapid mode. Hybridization mixes were prepared with 0.375 fmol PCR-amplified library that had been denatured with standard NaOH treatment and brought to a final volume of 125 µL with HT1 hybridization buffer (Illumina, 3 pM library in final mix). Following standard cluster generation and sequencing-primer hybridization, two dark cycles were performed for the splint-ligation libraries (i.e., two rounds of standard sequencing-by-synthesis in which imaging was skipped), which extended the sequencing primer by 2 nt, thereby enabling measurement of poly(A) tails terminating in non-adenosine bases. For the direct-ligation libraries, six dark cycles were performed instead of two, which extended the sequencing primer past the four random-sequence nucleotides in the 3′ adapter and then the last two residues of the tail.

Following the two dark cycles, a custom primer-extension reaction was performed on the sequencer using 50 µM dTTP as the only nucleoside triphosphate in the reaction. To perform this extension, the flow cell temperature was first set to 20°C. Then, 120 µL of universal sequencing buffer (USB, Illumina) was flowed over each lane, followed by 150 µL of Klenow buffer (NEB buffer 2 supplemented with 0.02% Tween-20). Reaction mix (Klenow buffer, 50 µM dTTP, and 0.1 U/µL Large Klenow Fragment, NEB) was then flowed on in two aliquots (150 µL and 100 µL). The flow-cell temperature was then increased to 37°C at a rate of 8.5°C per min and the incubation continued another 2 min after reaching 37°C. 150 µL of fresh reaction mix was then flowed in, and following a 2 min incubation, 75 μL of reaction mix was flowed in eight times, with each flow followed by a 2 min incubation. The reaction was stopped by decreasing the flow cell temperature to 20°C, flowing in 150 μL of quench buffer (Illumina HT2 buffer supplemented with 10 mM EDTA) and then washing with 75 μL of HT2 buffer. The flow cell was prepared for subsequent sequencing with a 150 μL and a 75 μL flow of HT1 buffer (Illumina). 50 cycles of standard sequencing-by-synthesis were then performed to yield the first sequencing read (read 1). XML files used for this protocol are provided at https://github.com/kslin/PAL-seq.

The flow cell was stripped, a barcode sequencing primer was annealed, and seven cycles of standard sequencing-by-synthesis were performed to read the barcode. The flow cell was then stripped again, and the same primer as used for read 1 was hybridized and used to prime 250 cycles of standard sequencing-by-synthesis to generate read 2. Thus, each PAL-seq tag consisted of three reads: read 1, read 2, and the indexing (barcode) read. For cases in which a tag corresponded to a polyadenylated mRNA, read 1 was the reverse complement of the 3′ end of the mRNA immediately 5′ of the poly(A) tail and was used to identify the mRNA and cleavage-and-polyadenylation site of long-tailed mRNAs. The indexing read was used to identify the sample, and read 2 was used to measure poly(A)-tail length and identify the mRNA and cleavage-and-polyadenylation site of short-tailed mRNAs. The intensity files of reads 1 and 2 were used for poly(A)-tail length determination, along with the Illumina fastq files.

PAL-seq v2 data analysis

Tail lengths for the splinted-ligation data were determined using a Gaussian hidden Markov model (GHMM) from the python2.7 package ghmm (http://ghmm.org/), analogous to the model used in TAIL-seq (Chang et al., 2014) and described in the next paragraph. Read 1 was mapped using STAR (v2.5.4b) run with the parameters '--alignIntronMax 1 --outFilterMultimapNmax 1 --outFilterMismatchNoverLmax 0.04 --outFilterIntronMotifs RemoveNoncanonicalUnannotated --outSJfilterReads', aligning to an index of the mouse genome built using mm10 transcript annotations that had been compressed to unique instances of each gene selecting the longest transcript and removing all overlapping transcripts on the same strand (Eichhorn et al., 2014). The genome index also included sequences of the quantification spikes and the common portion of the poly(A)-tail length standards. The sequences that identified each RNA standard (the last 20 nt of each standard sequence, Table S3) were not aligned using STAR. Instead, the unix program grep (v2.16) was used to determine which reads matched each standard (allowing no mismatches), and these reads were added to the aligned reads from the STAR output. Tags corresponding to annotated 3′ UTRs of mRNAs were identified using bedtools (v2.26.0), and if the poly(A)-tail read (read 2) contained a stretch of ≥ 10 T residues (the reverse complement of the tail) in an 11-nt window within the first 30 nt, this read was carried forward for GHMM analysis. If read 2 failed to satisfy this criterion but began with ≥ 4 T residues, the tail length was called based on the number of contiguous T residues at the start of read 2; by definition, these tails were < 10 nt and thus easily determined by direct sequencing.

For each read 2 that was to be input into the GHMM a 'T signal' was first calculated by normalizing the intensity of each channel for each cycle to the average intensity of that channel when reading that base in read 1 and then dividing the thymidine channel by the sum of the other three channels. Sometimes a position in a read would have a value of 0 for all four channels. A read was discarded if it contained more than five such positions. Otherwise, the values for these positions were imputed using the mean of the five non-zero signal values upstream and downstream (ten positions total) of the zero-valued position. A three-state GHMM was then used to decode the sequence of states that occurred in read 2. It consisted of an initiation state (state 1), a poly(A)-tail state (state 2), and a non-poly(A)-tail state (state 3). All reads start in state 1. From state 1 the model can remain in state 1 or transition to state 2. From state 2 the model can either remain in state 2 or transition to state 3. The model was initialized with the following transition probabilities:

$from \setminus to$	$state_1$	$state_2$	$state_3$
$state_1$	0.001	0.95	0.049
$state_2$	0.001	0.95	0.049
$state_3$	0.001	0.001	0.998

The initial emissions were Gaussian distributions with means of 100, 1, and −1 and variances of 1, 0.25 and 0.25, respectively. In general, the emission Gaussians for the model corresponded to the logarithm of the calculated T signal at each sequenced base in read 2. The initial state probabilities were 0.998, 0.001, and 0.001 for states 1, 2 and 3, respectively.

After initializing the model, unsupervised training was performed on 10,000 randomly selected PAL-seq tags, and then the trained model was used to decode all tags, with the number of state 2 cycles reporting the poly(A)-tail length for a tag. Only genes with ≥ 50 poly(A)-tail length measurements were considered for analyses involving mean poly(A)-tail lengths.

Analysis of PAL-seq data from the ss-ligation protocol

To account for mRNAs with very short tails or extensive terminal modifications, we implemented a version of PAL-seq that did not use splinted ligation. Tail lengths from these ss-ligation datasets, acquired for steady-state samples from both cell lines, were determined using a modified version of the PAL-seq analysis pipeline written for python3. The T-signal in this pipeline was modified to allow more accurate quantification of 0-length tails. Instead of normalizing the intensity of each channel for each cycle to the average intensity of that channel when reading that base in read 1, the intensity of each channel was normalized to the average intensity of the channels for the other three bases in read 1. The intensity of the T channel was then divided by the sum of the other channel intensities to calculate the T signal, and tails were called using the hmmlearn package (v0.2.0). Tags representing short tails, including short tails that ended with many non-A residues, were identified as those for which read 1 and read 2 mapped to the same mRNA 3′ UTR (usually ~4% of the tags). Tail lengths for these tags were called without the use of the GHMM. Instead, their tail lengths were determined by string matching, allowing any number of untemplated U residues but no more than two G or C residues to precede the A stretch. Tags not identified as representing short-tails were analyzed using the GHMM, excluding from further analysis occasional outliers determined by the GHMM to have tails ≤ 8 nt.

Most of the tags that had either only a very short tail or no tail did not correspond to mRNA cleavage-and-polyadenylation sites. Therefore, to be carried forward in our analysis, short-tailed tags were required to have a 3′-most genome mapping position (as determined from read 1 but requiring that read 2 also map uniquely to the same 3′ UTR) that fell within a 10 nt window of a PAL-seq–annotated cleavage-and-polyadenylation site.

Although the single-stranded ligation protocol provided the opportunity to account for mRNAs with very short or highly modified tails, examination of the recovery of internal standards indicated that tags representing longer tails (≥ 100 nt) were not as well recovered in the datasets in which we implemented ss ligation. Therefore, for steady-state samples from each cell line, we generated composite tail-length distributions in which the ss-ligation dataset contributed to the distribution of tails < 50 nt, and the splint-ligation dataset contributed to the distribution of tails ≥ 50 nt. For example, *Slc38a2* had 635 standard PAL-seq tags, 169 of which (~27%) had tails < 50 nt, and this same gene had 703 ss-ligation PAL-seq tags, 393 of which (~56%) had tails < 50 nt. The composite tail-length distribution replaced the 169 short-tailed splint-ligation PAL-seq tags with the 393 short-tailed ss-ligation PAL-seq tags, normalizing the latter cohort by a scaling factor. This scaling factor was determined from the ratio of the counts of the splint-ligation tags with tail lengths between 30–70 nt (135 tags) to the counts of the corresponding tags in the ss-ligation dataset (153 tags).

3′-end annotations were generated from PAL-seq tags with tails ≥ 11 nt, using an algorithm previously developed for data from poly(A)-position profiling by sequencing (3P-seq) (Jan et al., 2011). Each PAL-seq read 1 that mapped (with at least 1 nt of overlap) to an annotated 3′ UTR (Eichhorn et al., 2014) was compiled by the genomic coordinate of its 3′-most nucleotide. The position with the most mapped reads was annotated as a 3′ end. All reads within 10 nt of this end (a 21 nt window) were assigned to this end and removed from subsequent consideration. This process was repeated until there were no remaining 3′ UTR-mapped reads. For each gene, the 3′-end annotations were used in subsequent analyses if they accounted for ≥ 10% of the 3′ UTR-mapping reads for that gene.

Documentation and code to calculate and analyze T signals and determine tail lengths are available for both the splint-ligation and ss-ligation pipelines at https://github.com/kslin/PAL-seq.

TAIL-seq library preparation, sequencing, and analysis

The 2 h time-interval TAIL-seq used for comparison with PAL-seq was prepared using the same library cDNA as was used for PAL-seq v2 libraries, but amplifying the library using different primers (Table S3). Amplification and purification were as for PAL-seq v2. Samples were sequenced with either a paired-end 50-by-250 run (2 h time-interval sample) using a HiSeq 2500 operating in normal mode using a v3 kit. Other Illumina sequencing chemistries (including v1, v2, and v4 kits run in rapid and normal modes) did not yield accurate tail-length measurements when used in paired-end mode. Analysis was as described for PAL-seq v2, except a five-state GHMM was used (Chang et al., 2014) to accommodate the difference in the nature of the T-signal output imparted by the different mode of sequencing. The five states were an initiation state, a poly(A) state, a poly(A) transition state, a non-poly(A) transition state, and a non-poly(A) state.

RNA-seq

Fragmented poly(A)-selected RNAs were supplemented with three short quantification standards (Table S3), and then ligated to adapters, reverse-transcribed, and amplified to prepare the RNA-seq and ribosome-profiling libraries, respectively (Subtelny et al., 2014). These libraries were sequenced on an Illumina HiSeq 2500. For all RNA-seq data, only reads mapping to ORFs of annotated gene models (Eichhorn et al., 2014) were considered, excluding the first 50 nt of each ORF, which was implemented to match ribosome-profiling data of a contemporary study examining the effects of miRNAs (Eisen et al., 2019). A cutoff of ≥ 10 reads per million mapped reads (RPM) was applied to each sample.

Calculation of mRNA half-lives

Half-lives were estimated independently from both RNA-seq data and PAL-seq tag abundance. Prior to half-life fitting, mRNA abundances were normalized across time intervals based on the quantification standards added to each sample prior to library preparation.

Half-lives were determined by fitting to the equation

$$m(t_i) = \left(\frac{\alpha}{\beta}\right)(1 - e^{-\beta(t_i - t_{off})})(\delta) \tag{1}$$

in the case of the continuous-labeling experiment, or to the equation

$$m(t_i) = \frac{\alpha}{\beta}e^{-\beta t_i} + c \tag{2}$$

in the case of the transcriptional shutoff experiment, where $m(t_i)$ is the expression of an mRNA at a given time i, α is the rate of mRNA production, β is the rate of mRNA degradation, t_{off} is a global time offset, δ is a global scaling parameter to adjust the steady-state time point, and c is a baseline for the final expression of the gene in the transcriptional-shutoff experiment. Because the quantification standards were not applicable to the steady-state sample, the steady-state sample was normalized by a globally-fitted constant (setting t_i to 100 h for this time interval).

Because the half-life fitting for the continuous-labeling experiment required the global parameters t_{off} and δ, half-lives for all genes needed to be fit simultaneously. Accordingly, we minimized the least-square errors loss function (L_2).

$$L_2(p) = \sum_{i}^{I}\sum_{j}^{J}(\ln(x_{ij}(p)) - \ln(data_{ij}))^2 , \tag{3}$$

for the simulated number of normalized tags x at time point i for gene j. The total number of time points and genes are denoted by I and J, respectively. L_2 depends on the parameters p ($\alpha_{i,j}$, $\beta_{i,j}$, t_{off}, and δ). The optimization for $\alpha_{i,j}$, $\beta_{i,j}$, t_{off}, and δ was performed using the L-BFGS-B method in the optim function in R.

To increase the efficiency of the optimization, we also implemented an analytical gradient for this model. This gradient computed the quantity $\frac{dL_2}{dp}$ which, when passed to the optimizer, decreased the number of iterations required to minimize the loss. This quantity was computed for each of the parameters as follows

$$\frac{dL_2}{d\alpha_j} = \sum_i^I (\zeta \left(\frac{1}{\beta_j}\right)(1 - e^{\left(-\beta_j(t_i - t_{off})\right)}))(\delta) \tag{4.1}$$

$$\frac{dL_2}{d\beta_j} = \sum_i^I (\zeta \left(\frac{\alpha_j}{{\beta_j}^2}\right) e^{\left(-\beta_j(t_i - t_{off})\right)} - \left(\frac{\alpha_j}{{\beta_j}^2}\right) + \left(\frac{\alpha_j}{\beta_j}\right) t_i e^{\left(-\beta_j(t_i - t_{off})\right)} - \left(\frac{\alpha_j}{\beta_j}\right) t_{off} e^{\left(-\beta_j(t_i - t_{off})\right)})(\delta) \tag{4.2}$$

$$\frac{dL_2}{dt_{off}} = \sum_i^I \sum_j^J (\zeta(-\alpha_j) e^{\left(-\beta_j(t_i - t_{off})\right)})\ (\delta), \tag{4.3}$$

$$\frac{dL_2}{d\delta} = \sum_i^I \sum_j^J (\zeta \left(\frac{\alpha_j}{\beta_j}\right)(1 - e^{\left(-\beta_j(t_i - t_{off})\right)}))\ , \tag{4.4}$$

where ζ is the first component of the derivative of the loss function

$$\zeta = 2\left(\ln\left(\left(\frac{\alpha_j}{\beta_j}\right)\left(1 - e^{\left(-\beta_j(t_i - t_{off})\right)}\right)\delta\right) - \ln\,(data_{ij})\right)\frac{\beta_j}{\delta\alpha_j(1 - e^{\left(-\beta_j(t_i - t_{off})\right)})}\ , \tag{5.1}$$

and δ is further defined by the piecewise function

$$\delta \ = \begin{cases} \delta_p \text{ for } t_i \ = \ t_{ss} \\ 1 \text{ otherwise} \end{cases} . \tag{5.2}$$

Ranges of rates constants fit to this exponential model and the subsequent deadenylation model were truncated to reflect the lack of confidence in values of and differences between extreme outliers. Half-live values were truncated to fall between 6 min and 100 h, deadenylation rate constants were truncated to fall between 0.03 and 30 nt/min, decapping rate constants at 20 nt were truncated to fall between 0.003 and 3 min^{-1}, and production rate constants were truncated to fall between 10^{-8} and 10^{-5} min^{-1}. All calculations (including correlations) were performed on the non-truncated values.

Model of mRNA production, deadenylation, and decay

The model of mRNA production, deadenylation, and decapping (decay) was a system of differential equations

$$\frac{dA_n}{dt} = k_0 - (k_1 + k_2)(A_n) \tag{6.1}$$

$$\frac{dA_{n-1}}{dt} = k_0 + k_1(A_n) - (k_1 + k_2)A_{n-1} \tag{6.2}$$

$$\frac{dA_{n-2}}{dt} = k_0 + k_1(A_{n-1}) - (k_1 + k_2)A_{n-2} \tag{6.3}$$

$$\vdots$$

$$\frac{d(A_0)}{dt} = k_0 + k_1(A_1) - k_2(A_0)\,, \tag{6.4}$$

where A_n is an mRNA with tail length n, and k_0, k_1 and k_2 are rate constants that describe the production, deadenylation and decapping rates, respectively. The final deadenylation product (A_0) has a deadenylation rate constant of zero, as it has no tail. The rate constants k_0 and k_2 are themselves functions of tail length (l), specified by the respective negative binomial and logistic functions

$$k_0(l) = \frac{\alpha\Gamma(v_p + l)}{l!\,\Gamma(v_p)}\left(\frac{m_p}{v_p + m_p}\right)^l\left(\frac{v_p}{v_p + m_p}\right)^{v_p} \tag{7.1}$$

$$k_2(l) = \frac{\beta}{\left(1+e^{-\frac{l-m_d}{v_d}}\right)} , \tag{7.2}$$

where the parameters α, β, v_p, m_p, m_d, v_d, are fitted parameters. The parameters α and β are scaling terms for production and decapping distributions, respectively. The parameters m_p and m_d describe the expected value of those distributions, and v_p and v_d describe the spread.

Equations (6) were re-written as a linear, time-invariant (LTI) system (Dahleh et al., 2004)

$$\dot{x} = \begin{bmatrix} -k_1 - k_2(l_n) & 0 & \cdots & 0 & 0 \\ k_1 & -k_1 - k_2(l_{n-1}) & & & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & & -k_1 - k_2(l_1) & 0 \\ 0 & 0 & \cdots & k_1 & -k_2(l_0) \end{bmatrix} x(t) + \begin{bmatrix} k_0(l_n) \\ k_0(l_{n-1}) \\ \vdots \\ k_0(l_1) \\ k_0(l_0) \end{bmatrix} , \tag{8}$$

or, more succinctly, as

$$\dot{x} = Cx(t) + D \ , \tag{9}$$

where the coefficient matrix, C, is specified by the coefficients of the differential equations (6), and the source vector D is specified by the production rate. C is a bidiagonal 251×251 matrix, whereas D, $x(t)$, and $\dot{x}$ are 251×1 vectors. In the case of the continuous-labeling experiment, $x(t = 0) = \mathbf{0}$. The transcriptional-shutoff experiment begins with $x(t = -1 \text{ h}) = \mathbf{0}$, but $x(t = 0)$ is determined by the values of the system after 1 h of simulation.

Equation (9) has the analytical solution

$$x(t) = C(e^{Ct} - I)D \ , \tag{10}$$

where I is the identity matrix and e^{Ct} is the matrix exponential of the coefficient matrix scaled by time. Both this analytical solution and numerical integrators (which do not require an analytical solution) can be used to compute the result. We found that numerical stability and computational efficiency were optimal when using the LSODE solver with parameters set for a banded Jacobian matrix in the deSolve package (v1.21) of R, with the model written in C and dynamically loaded into R. Increasing the number of allowed tail-length states from 250 to 300 had little effect on the resulting fitted rate constants but greatly increased computation time.

The model yielded abundances for each tail-length isoform at each time interval, using seven parameters for each gene, three of which were shared across all of the genes (Table S2). From these abundances, the residual sum of squares was computed from the corresponding standard-normalized PAL-seq datasets. Although tail lengths ≥ 250 nt were modeled, measurements for these lengths were not available from PAL-seq v2, and thus were excluded from the fitting. Likewise, tail-lengths < 20 were modeled for all time intervals, but because the abundance of tail lengths < 20 nt was only available for steady state, these lengths were excluded from fitting all but the steady-state interval. As a result, for the continuous-labeling experiment using cell line 1, parameters for each gene were fit to 1400 data points (230 tail lengths × 5 time intervals + 250 for the steady-state), and for the experiment using the cell line 2, parameters for each gene were fit to 1170 data points. The optimization was performed using the L-BFGS-B method in the optim function of R, or, in the case of the global fitting, using the L-BFGS-B method in the NLopt package (v1.0.4) of R.

A simple L_2 loss function skewed the fits to the time intervals that had larger values. A common solution to this problem is to fit to log-transformed values, but because our data were sparse, with many tail-length positions having zero tags, pseudo-counting to allow log-space fitting resulted in poor fits. Therefore, residuals were variance weighted using the loss function

$$L_2(p) = \sum_{i}^{I}\sum_{j}^{J}\sum_{k}^{K}\left(\frac{(x_{ijk}(p) - data_{ijk})^2}{\mathrm{Var}(data_i)}\right), \tag{11}$$

where i, j, and k are the time-interval, gene, and tail length with I, J and K as the maximal values of time-intervals, genes, and tail lengths. The variance of a dataset at a single time-interval is given by $\mathrm{Var}(data_i)$.

The model is constrained by the 0-tail-length species, which builds up when decapping is slow with respect to deadenylation. Such a buildup was observed in the steady-state tail-length distribution of short-lived mRNAs but occurred primarily between 0 and 20 nucleotides (Figure 6C). Because of this discrepancy, a composite residual was calculated for the model and the data. Abundances for tails < 20 nt were averaged and this average was used to replace the abundances for each tail length < 20 for the steady-state data. In addition, when comparing the associated short tails from the data and the model, the residuals for tails < 20 were weighted by either 6- or 5-fold (cell lines 1 and 2, respectively) to account for opting not to fit to measurements for tails < 20 nt in the non-steady–state samples.

As with gene-specific parameters, global parameters v_p, m_d, and v_d were fit using pre-steady-state measurements of tails ranging from 20–249 nt. The composite steady-state tail-length distributions of Figure 2A were also used, which constrained buildup of short-tailed mRNAs. Fitting was performed on subsets of 100 genes (selected randomly without replacement from genes with composite steady-state distributions, yielding 22 and 15 subsets for cell lines 1 and 2, respectively), including v_p, m_d, and v_d in the parameter vector. Median values of the global parameters (Table S1) were then used to fit each gene-specific parameter.

Bootstrap analysis

Tags in the cell line 1 PAL-seq dataset were resampled 10 times with replacement and assigned to a gene and tail length based on a multinomial probability distribution generated from the counts for each tail length in the original dataset. These resampled datasets were then used for background subtraction, global parameter determination, and model fitting.

Background subtraction and normalization for PAL-seq data

Although the efficacy of the 5EU purification enabled efficient enrichment of labeled RNAs at short time intervals (Eisen et al., 2019), we also modeled and corrected for residual background caused by non-specific binding of the unlabeled RNA to the streptavidin beads (Figure S2F).

We designed our background model under the assumption that the background in the time courses stems primarily from the capture of a fixed amount of non-5EU labeled mRNA during the 5EU purification. Accordingly, we subtracted a fraction (0.3%) of the steady-state data from each continuous-labeling dataset. This fraction of input sample was chosen such that at 40 min long-lived genes (half-life $\geq$ 8 h) had no mRNAs with tail lengths ≤ ~100 nt on average, but short-lived genes (half-life ≤ 30 min) were unaffected (Figure S2F). Likewise, we subtracted standard-normalized time-interval–matched input data from each transcriptional-inhibition dataset, as actD influenced which unlabeled cellular mRNAs were available to contribute to the background. The fraction of each input sample to subtract was chosen such that at 0 h long-lived genes (half-life ≥ 8 h) had no mRNAs with tail lengths ≤ ~100 nt on average, but short-lived genes (half-life ≤ 30 min) were unaffected. Genes were included in the final background-subtracted set only if the sum of their background-subtracted tag counts was ≥ 50 tags.

After background subtraction, PAL-seq datasets were scaled to each time interval by matching the total number of background-subtracted tags for all genes at all tail lengths to the total number of tags for all genes for the corresponding time interval in the RNA-seq data. The scaled PAL-seq data were then used to compute half-lives for each gene, scaling the steady-state sample using a globally fitted constant.

ActD treatment

Cell line 2 was cultured as in the continuous-labeling experiments. We prepared 2, 2, 2, 3, and 4 500 cm^2 plates for the 0, 1, 3, 7 and 15 h time intervals, respectively. 5EU (400 µM final) was added to each plate (with one non-5EU plate prepared in parallel), and after 1 h actD (5 µg/mL final concentration, Sigma-Aldrich) was added. Cells were harvested as described for the continuous-labeling experiments, except that a quantitative spike RNA containing 5EU and corresponding to the chloramphenicol-resistance gene sequence (Table S3) was added to the lysis buffer at a concentration of 0.57 ng/mL, or 2 ng/plate. This RNA was prepared using an in vitro transcription reaction as above, with a 5EUTP-to-UTP ratio of 1:20.

Accession numbers

Raw and processed RNA-seq, PAL-seq, and TAIL-seq data is available at the GEO, accession number GSE134660. Code for configuring an Illumina HiSeq 2500 machine for PAL-seq and for calculation of tail lengths from PAL-seq or TAIL-seq data is available at https://github.com/kslin/PAL-seq.

Figure Legends

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 either miR-155 or miR-1 (cell lines 1 and 2, respectively) but cultured without microRNA induction. The 8 is in parenthesis 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 RNA 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 ≥ 250 nt are not shown.

(C) Distibutions 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 = 3048). Each bin is 2 nt. Genes with mean mRNA tail-length values greater than ≥ 250 nt were assigned to the 250 nt bin.

(D) Tail lengths over time. Mean tail lengths for mRNAs from each gene (n = 3048) are plotted along with box-and-whiskers overlays (line, median; box, 25th to 75th percentiles; whiskers, 5th to 95th percentiles). See also Figures S1 and S2.

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–D, S3G.

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 represented as a heatmap (bottom), with the range of coloration corresponding to the 5–95 percentile 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 now shown.

See also Figures S3F, S3H–J.

Figure 4. Computational Model of mRNA Deadenylation and Decay Dynamics

(A) Schematic of the computational model. Poly(A)-tail lengths of mRNAs are represented by A_n, where n is the length of the poly(A) tail. k_0, k_1, and k_2 are terms for mRNA production, deadenylation, and decapping, respectively, and Ø represents the loss of the mRNA molecule. The curves (right) indicate the distributions used to model probabilities of production and decapping 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 (left curve), whereas the decapping rate constant, β, scales the decapping distribution (right curve) (Table S2).

(B) Correspondence between the fit of 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, $n = 2778$). 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.

(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 as a function of 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–B, and Tables S1–S2.

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 = 2778).

(B) Tail lengths at which mRNAs are decapped, 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 decapping rate constant k_2 to yield a distribution of decapping events over all tail lengths. Plotted is the combined distribution for all mRNA molecules of all 2778 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 decapping rates of these species.

(C) Mean tail lengths at which mRNAs from each gene (n = 2778) were decapped, as inferred by the model. Otherwise, as in (B).

(D) Distribution of decapping 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 = 2778).

(E) Correlation between the deadenylation rate constant (k_1) and the decapping rate constant (k_2) at a tail length of 20 nt. The dashed line indicates $y = x$.

See also Figure S4.

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 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 Figure S5C–J.

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 RNA 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 = 2155). 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 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.

Supplemental Figure Legends

Figure S1. PAL-seq v2 Methodology and Benchmarking, Related to Figure 1

(A) Schematic of PAL-seq v2. The original version of PAL-seq (Subtelny et al., 2014) was modified to include an additional splint oligonucleotide capable of ligating to tails with a terminal U (step 1); two dark cycles prior to the primer-extension reaction (step 13), which prevented non-adenosine terminal residues from terminating the subsequent primer extension; primer extension through the tail with dTTP as the only nucleoside triphosphate (step 14); sequencing on a HiSeq machine, with the opportunity for multiplexing (steps 16 and 17); an additional read using the read 1 sequencing primer (read 2), which collected sequence and intensity information used to call poly(A)-tail lengths, as in TAIL-seq (Chang et al., 2014) (steps 18 and 19).

(B) Recovery of RNA standards. Before preparing libraries, two sets of RNA standards were added to each of the 34 RNA samples analyzed by PAL-seq v2 in this study and an accompanying study (Eisen et al., 2019). Set 1 contained seven RNAs with different tail lengths, and set 2 contained four RNAs with different tail lengths (Table S3). For each set of standards, the relative abundance of each standard in the final sequencing output was compared its relative abundance in the initial standard mixture, and this recovery ratio is plotted for each sample on a log scale. The relative recovery of standards varied somewhat, with no systematic bias that would indicate substantial depletion of poly(A)-tails of certain lengths. The 30 nt standard from set 2 was excluded from this analysis because it is an equal mixture of four different standards that end in a terminal A, C, G or U (Table S3), which was added to assess the ability to detect tails with a terminal U, as described in the next panel.

(C) Terminal nucleotide compositions of RNAs with tail measurements ≥ 5 nt. Libraries were prepared using a 5:1 mixture of splint oligos that would hybridize perfectly to either the 3′ end of RNAs ending in eight adenosines or the 3′ end of RNAs ending in seven adenosines followed by a terminal uridine, respectively. Left: Terminal nucleotide composition of PAL-seq v2 tags from the RNA standards for which poly(A) tails were prepared using poly(A) polymerase and ATP. These standards were expected to terminate exclusively with adenosine. Middle: Terminal nucleotide composition of PAL-seq v2 tags from the synthetic 30 nt standard, which was prepared with a tail designed to have an equal mixture of terminal A, C, G, or U. Although the splint oligonucleotides perfectly matched the versions ending A and U, the terminal U was somewhat depleted compared to the terminal A, and a substantial fraction of terminal G was also captured, perhaps due to wobble-pairing between the T in the splint and the terminal G in the standard. Right: Terminal nucleotide composition of PAL-seq v2 tags from mRNAs.

(D) The tail length distributions of the synthetic RNA standards, as measured by PAL-seq v2. Plotted is the cumulative distribution of poly(A)-tail lengths for each standard in the steady-state sample from cell line 1. The poly(A)-tail lengths measured by polyacrylamide gel electrophoresis (Subtelny et al., 2014) are indicated (key).

(E) Mean poly(A)-tail lengths of the two sets of synthetic standards, as measured by PAL-seq v2. For each standard, the mean tail length in each of the 34 samples in this study and an accompanying study (Eisen et al., 2019), as measured using PAL-seq v2, is plotted (black points). Also plotted is the mean tail length of each standard, as determined using denaturing gels (red crosses). Tails that exceeded 250 nt were not expected to be measured accurately, because their length exceeded the length of the sequencing read used to measure the tail.

(F) Comparison of the frequencies in which UTR 3′ ends were called by biological replicates of PAL-seq v2 (left) or the two different versions of PAL-seq (right). Each point is the number of tags that mapped to a genomic position, adding a pseudo-count of 0.1 tags. Dashed lines represent equivalency, after accounting for different sequencing depths.

Figure S2. Reproducibility of PAL-seq v2, Related to Figure 1

(A) Comparisons of biological replicates for different library preparation and tail-profiling protocols. For each gene that passed a 50-tag cutoff, mean poly(A)-tail lengths after 2 h of continuous labeling are shown for PAL-seq (left panel), our implementation of TAIL-seq (Lim et al., 2016), (middle panel), or PAL-seq v2 (right panel). Whereas the RNA examined using TAIL-seq and PAL-seq v2 datasets was isolated using 5EU labeling, the RNA examined using PAL-seq v1 dataset was isolated using 4-thiouridine labeling. The 2 h time interval was chosen for this analysis because its broad range of average tail lengths made it most suitable for comparing the results of different methods (Figure 1C). The dashed line represents y = x.

(B) Comparisons between different tail-profiling protocols. Compared are mean tail lengths generated by TAIL-seq and PAL-seq v2 (left panel), PAL-seq v1 and PAL-seq v2 (middle panel), and TAIL-seq and PAL-seq v1 (right panel). Otherwise, as in (A).

(C) Recovery of tail standards in the PAL-seq v1, TAIL-seq (splinted ligation) and PAL-seq v2 steady-state (single-stranded ligation) datasets. For analyses of the recovery of tail standards in PAL-seq v2 (splint-ligation) datasets, see Figure S1B. All six libraries contained seven standards from standard set 1; the TAIL-seq libraries contained three standards from standard set 2; and the PAL-seq v1 library contained seven standards from standard set 3. Tail lengths of the standards as determined by polyacrylamide-gel electrophoresis (Subtelny et al., 2014) are indicated (key) and shown as x-axis labels. The 30 nt standard from set 2 was excluded from this analysis because it was an equal mixture of four different standards that ended in A, C, G or U (Table S3) and was added to assess the ability to detect tails with a terminal U. The relative abundances of the standards in the sequencing data were quantified and compared to their relative starting abundance, and this recovery ratio is plotted for all samples. The values of each library were normalized to the abundance of the 107 nt standard in set 1.

(D) Mean tail lengths of the standards shown in (C) and the 30 nt standard. Otherwise, as in (C). For analysis of mean tail lengths of the standards in PAL-seq v2 (splint-ligation) datasets, see Figure S1E.

(E) Uridylation frequency as a function of tail length. The fraction of single uridine residues at the 3′ end of mRNA-mapping tags is plotted as a function of tail lengths ≥ 5 nt (black) along with a LOESS smoothing kernel (blue, with 5th–95th percent confidence intervals in grey) for either cell line 2 (top panels) or cell line 1 (bottom panels). These values were scaled by a factor of 5.23 to correct for the depletion of tags containing a terminal uridine, estimated from the ratio of tags mapping to the 30 nt standards terminating with either A or U, which had been added to the libraries at an equal molar ratio (Figure S1C). Uridine fractions corresponding to tail lengths ≥ 246 nt were combined into one bin at 246 nt.

(F) Effects of background subtraction of PAL-seq data at the earliest (40 min) time interval. Distributions of the unsubtracted (blue) and background-subtracted (red) tail lengths for short-lived (half-life < 30 min, n = 293 for both unsubtracted and background subtracted) and long-lived (half-life > 8 h, n = 379) mRNAs. The background subtraction differentially affected the long half-life mRNAs, as these had a proportionally smaller amount of labeled relative to unlabeled RNA at short time intervals, and thus unlabeled RNAs contributed a larger fraction of their reads at these intervals.

Figure S3. Half-life and Initial Tail-Length Measurements, Related to Figures 2 and 3

(A) Pairwise correlations (R_s) of half-life measurements. n = 4485, 4748, 3048, 1743, and 4658 genes for cell line 1 poly(A)-selected, cell line 2 poly(A)-selected, cell line 1 PAL-seq, cell line 2 PAL-seq, and Schwanhäusser et al. 2011 samples, respectively.

(B) Distributions of half-lives for mRNAs from all genes (n = 3048), ribosomal-protein genes (RPGs, n = 31), or immediate-early genes (IEGs, n = 19) (Tullai et al., 2007) obtained using PAL-seq data from cell line 1.

(C) Distribution of mRNA half-lives obtained using PAL-seq data from cell line 1 (n = 3048).

(D) Comparison of published half-life measurements (Schwanhäusser et al., 2011) with those obtained from 5EU continuous labeling. Dashed line is $y = x$.

(E) Comparison of half-life measurements from the transcriptional-shutoff experiment and those obtained from the continuous-labeling experiment. Dashed line is $y = x$.

(F) Comparison of mean poly(A)-tail lengths of mRNAs isolated from cell line 2 after 40 min of labeling with those isolated from cell line 1 after 40 min of labeling. Dashed line is $y = x$.

(G) Relationship between half-life and mean steady-state tail length of mRNAs in 3T3 cells. Tail lengths and half-lives were determined using only standard PAL-seq data, in which the adapter oligo was appended to the tail using splinted ligation. Otherwise, as in Figure 2A.

(H) The distribution of c.v. values of tail lengths after 40 min of labeling for mRNAs from each gene. Each c.v. value is the average of two biological replicates.

(I) Comparison of c.v. values of tail lengths after 40 min of labeling between two biological replicates.

(J) Relationship between mRNA half-life and c.v. values of tail lengths after 40 min of labeling. Each c.v. value is the average of two biological replicates.

Figure S4. Model Development and Testing, Related to Figures 4 and 5

(A) Schematic of the model with two deadenylation rates. Deadenylation is parameterized with two rate constants, one that describes deadenylation of tail lengths > 110 nt (k_1) and one that describes deadenylation of tails ≤ 110 nt (k_1'). The transition between these rates is determined by a generalized logistic function with a transition parameter arbitrarily set to 1 (a sharp transition). Otherwise, as in Figure 4A.

(B) Comparison of the residual sum of squares (RSS) between the model with two deadenylation rates (A) and the model with one deadenylation rate. Dashed line is $y = x$.

(C) Relationship between the second (k_1') and the first (k_1) deadenylation rate constant fit for mRNAs of each gene using the model in (A). Dashed line is $y = x$.

(D) Comparison of a model with two deadenylation rates in which the transition between the rates occurs at a tail length of 150 nt and the model with one deadenylation rate. Otherwise, as in (B).

(E) Relationship between the second (k_1') and the first (k_1) deadenylation rate fit for mRNAs of each gene using the model in (D). Dashed line is $y = x$.

(F) Distribution of R^2 values for all genes fit by the model (n = 2778). Dashed lines indicate the R^2 values of the four genes shown in Figure 4B.

(G) Analysis of the robustness of fitted rate constants to input parameter identities. The distributions of s.d. values of rate constants for all fitted genes over 10 rounds of fitting with varying input parameters are displayed as empirical cumulative distributions. The input parameters were randomly selected from a uniform distribution bounded by the 10th to 90th percentiles of rate constants of all genes during a previous round of fitting. Using an unbounded randomized parameter selection resulted in larger variation but also larger final residuals. The s.d. values for 90% of genes were less than 3.7×10^{-5}, 3.5×10^{-5}, 5.8×10^{-5}, and 2.4×10^{-4} for rate constants for starting tail length, production, deadenylation, and decapping, respectively (with all parameters shown as s.d. of the $\log_{10}$ of the value).

(H) Bootstrapping analysis of fitted rate constants. For each dataset, the total number of tags was resampled ten times based on a multinomial probability distribution specified by the original tag counts for every tail length position for each gene. These resampled datasets were then background subtracted and fit to the computational model. Shown for each fitted gene-specific parameter is the cumulative distributions of its c.v. values for all fitted genes. The s.d. for 90% of genes were less than 0.01 , 0.03, 0.06, and 0.49, for rate constants for starting tail length, production, deadenylation, and decapping, respectively (with all parameters shown as s.d. of the $\log_{10}$ of the value). The greater variation observed for the decapping parameter reflected the relatively few data points effectively used for its fitting, as this parameter related primarily to the short-tailed region of the distribution.

(I) Relationship between production rate from an exponential fit to the data and the production rate as determined by the computational model. Dashed line is $y = x$.

(J) Model reproducibility across biological replicates. Plots show the relationship between mean starting tail lengths (m_p, left panel), production-rate scaling terms (α, middle left panel), deadenylation rate constants (δ, middle right panel) or decapping-rate scaling terms (β, right panel) for mRNAs from the same genes in the two cell lines. Dashed line is $y = x$.

Figure S5. Analyses of Deadenylation Rate Constants and Steady-State Tail Lengths and Modifications, Related to Figures 5 and 6

(A) Relationship between mRNA half-life, as determined by an exponential fit to abundance, and deadenylation rate constant, as determined by the model. Because the structure of the computational model enhanced the correspondence between half-life and deadenylation rate constants, analyses performed with primary data (Figure 2B and Figure 7D) provide a more accurate indication of the correspondence between mRNA half-life and deadenlylation rate.

(B) Relationship between measured mRNA lifetime and mRNA lifetime inferred by the model. For mRNAs from each gene, the number of tail nucleotides separating the mean starting tail length (Figure S4J) and the mean tail length at decapping (Figure 5C) was multiplied by the deadenylation rate constant (Figure 5A). Measured lifetime (the inverse of the degradation rate from an exponential fit) was then compared to this inferred lifetime. The dashed line indicates y = x.

(C) Relationship between the steady-state abundance of short-tailed transcripts and mRNA half-life. For each tail length from 1–100 nt, the fraction of mRNAs with tail lengths that fell below that position was calculated for each gene, and the relationship (R_s) between this fraction falling below the tail-length cutoff and half-life was determined as in Figure 6A. These 100 R_s values are plotted as a function of the tail-length cutoff used to classify short-tailed transcripts. This analysis started with the composite steady-state tail-length distribution generated for the analysis of Figure 2A, which accounted for very short and highly modified tails. (n = 2778).

(D) The distribution of terminal uridylation on short- and long-lived mRNAs at steady state. The tail-length distributions of Figure 6B are replotted and colored by uridylation frequency (key).

(E) Examples of mRNA 3′-end isoforms plotted in Figure 6E. For the genomic locus corresponding to the dominant cleavage-and-polyadenylation site of *Actb*, several possible 3′-end isoforms lacking poly(A) tails are shown, along with distance from the dominant 3′ end and whether or not the isoform would be included in the plot of Figure 6E. Also depicted is a long-tail mRNA used to annotate the 3′ end of the UTR for the analyses of Figure 6E and Figures S6F–I.

(F) Distribution of tags as a function of the distance between the inferred 3′ end of their UTR and the dominant 3′ end. Only tags with a poly(A)-tail longer than 30 nt were used in this analysis.

(G) Fraction of tail-containing tags for each gene as a function of the distance between the inferred 3′ end of their UTR and the dominant 3′ end. As each dominant 3′ end was defined as the position represented by the most tags in a 21 nt window, in principle, no gene should have > 50% of its tags at a position other than the dominant end. However, because dominant 3′ end annotations were determined using data from a separate experiment (standard PAL-seq with splinted ligation to the mRNA 3′ end), > 50% of tags for some genes mapped to a position other than 0; these outliers represent discrepancies between biological replicates.

(H) Nucleotide composition near cleavage-and-polyadenylation sites. For each tag mapping to within 10 nt of an annotated 3′ end of an mRNA 3′ UTR, the frequency of each genomic nucleotide is plotted as a function of the distance from the annotated 3′ end. The depletion of A at position 0 and its enrichment at position 1 were artifacts of 3′-end annotation because any A at the final nucleotide of a 3′ UTR was assigned to the poly(A) tail, even if that A might have been genomically encoded.

(I) Relationship between the steady-state fraction of tails > 175 nt and mRNA half-life. Otherwise as in Figure 6A.

(J) Relationship between the steady-state abundance of long-tailed transcripts and mRNA half-life. For each tail length from 250–100, the fraction of mRNAs with tail lengths that fell above that position was calculated for each gene, and the relationship (R_s) between this fraction falling above the tail-length cutoff and half-life was determined as in (I) and plotted as in (C).

(K) Mean tail lengths for mRNAs from each gene plotted in Figure 6B. Violin plots show distributions of short- and long-lived mRNAs, with the median of each distribution shown as a horizontal line (and indicated above each group). Otherwise as in Figure 6B.

Table S1. Parameters of the Computational Model, Related to Figure 4 and Table 1.

Table of fitted parameters for the computational model. Staring tail lengths (m_p), production rate scaling terms (α), deadenylation rate constants (δ), and decapping rate scaling terms (β) were fit to the computational model. The global parameters v_p, m_d, and v_d were 16.27, 263.95, and 11.05 for cell line 1 and 15.5, 265.44, and 13.97 for cell line 2. mRNA half-lives were fit to an exponential model.

Figure 1

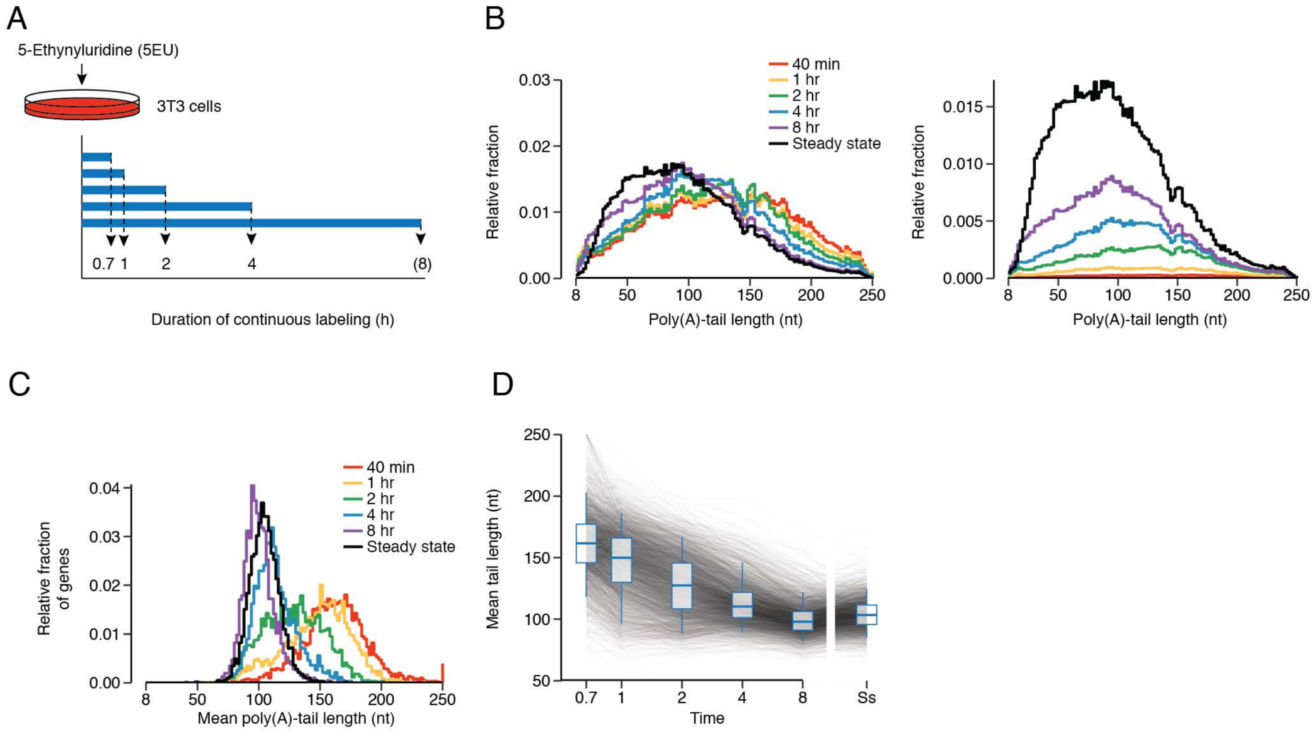

A
5-Ethynyluridine (5EU)
3T3 cells
0.7
1
2
4
(8)
Duration of continuous labeling (h)
B
40 min
1 hr
2 hr
4 hr
8 hr
Steady state
Relative fraction
0.03
0.02
0.01
0.00
8
50
100
150
200
250
Poly(A)-tail length (nt)
Relative fraction
0.015
0.010
0.005
0.000
8
50
100
150
200
250
Poly(A)-tail length (nt)
C
40 min
1 hr
2 hr
4 hr
8 hr
Steady state
Relative fraction of genes
0.04
0.03
0.02
0.01
0.00
8
50
100
150
200
250
Mean poly(A)-tail length (nt)
D
Mean tail length (nt)
250
200
150
100
50
0.7
1
2
4
8
Ss
Time

Figure 2

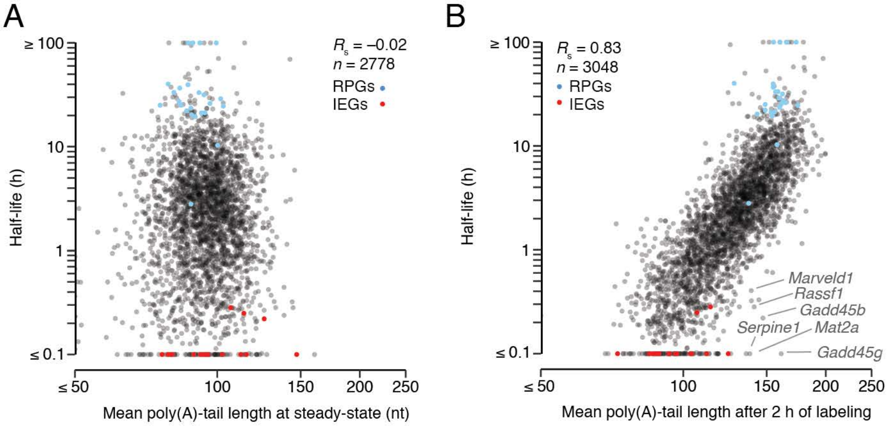

A
≥ 100
10
1
≤ 0.1
Half-life (h)
R_s = –0.02
n = 2778
RPGs ●
IEGs ●
≤ 50
100
150
200
250
Mean poly(A)-tail length at steady-state (nt)
B
≥ 100
10
1
≤ 0.1
Half-life (h)
R_s = 0.83
n = 3048
● RPGs
● IEGs
Marveld1
Rassf1
Gadd45b
Serpine1
Mat2a
Gadd45g
≤ 50
100
150
200
250
Mean poly(A)-tail length after 2 h of labeling

Figure 3

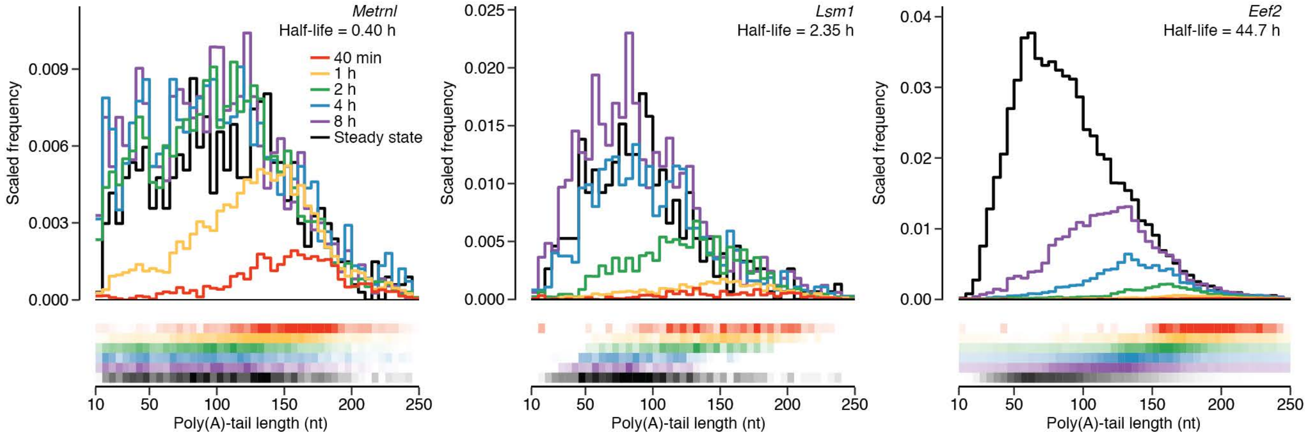

Metrnl
Half-life = 0.40 h
40 min
1 h
2 h
4 h
8 h
Steady state
Scaled frequency
0.009
0.006
0.003
0.000
Lsm1
Half-life = 2.35 h
0.025
0.020
0.015
0.010
0.005
0.000
Eef2
Half-life = 44.7 h
0.04
0.03
0.02
0.01
0.00
10
50
100
150
200
250
Poly(A)-tail length (nt)

Figure 4

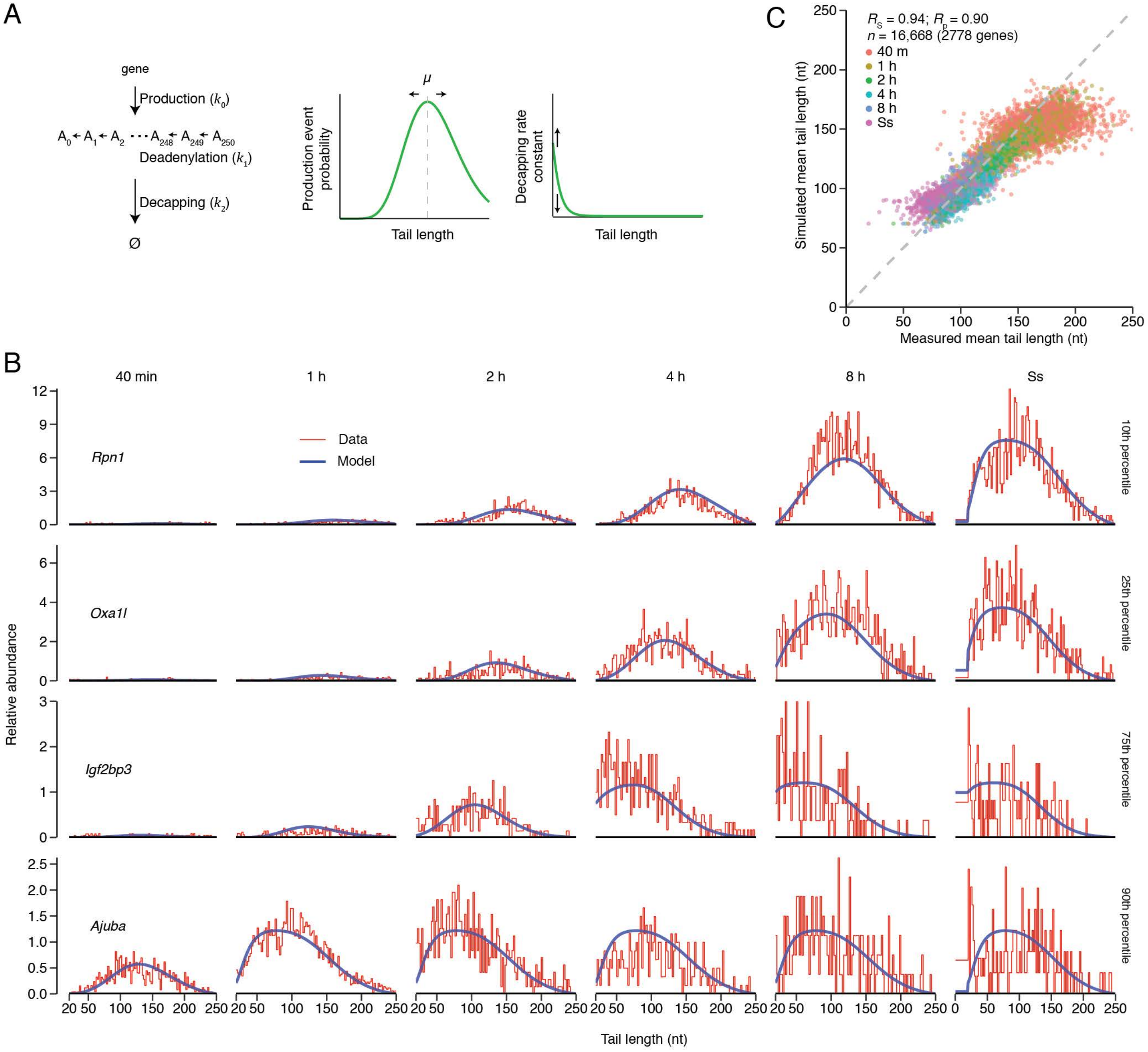

A
gene
Production (k_0)
A_0 ← A_1 ← A_2 ··· A_{248} ← A_{249} ← A_{250}
Deadenylation (k_1)
Decapping (k_2)
Ø
μ
Production event probability
Tail length
Decapping rate constant
Tail length
C
R_S = 0.94; R_p = 0.90
n = 16,668 (2778 genes)
40 m
1 h
2 h
4 h
8 h
Ss
Simulated mean tail length (nt)
Measured mean tail length (nt)
0
50
100
150
200
250
B
40 min
1 h
2 h
4 h
8 h
Ss
Data
Model
Rpn1
Oxa1l
Igf2bp3
Ajuba
10th percentile
25th percentile
75th percentile
90th percentile
Relative abundance
0
3
6
9
12
2
4
1
0.0
0.5
1.0
1.5
2.0
2.5
20 50 100 150 200 250
Tail length (nt)

Figure 5

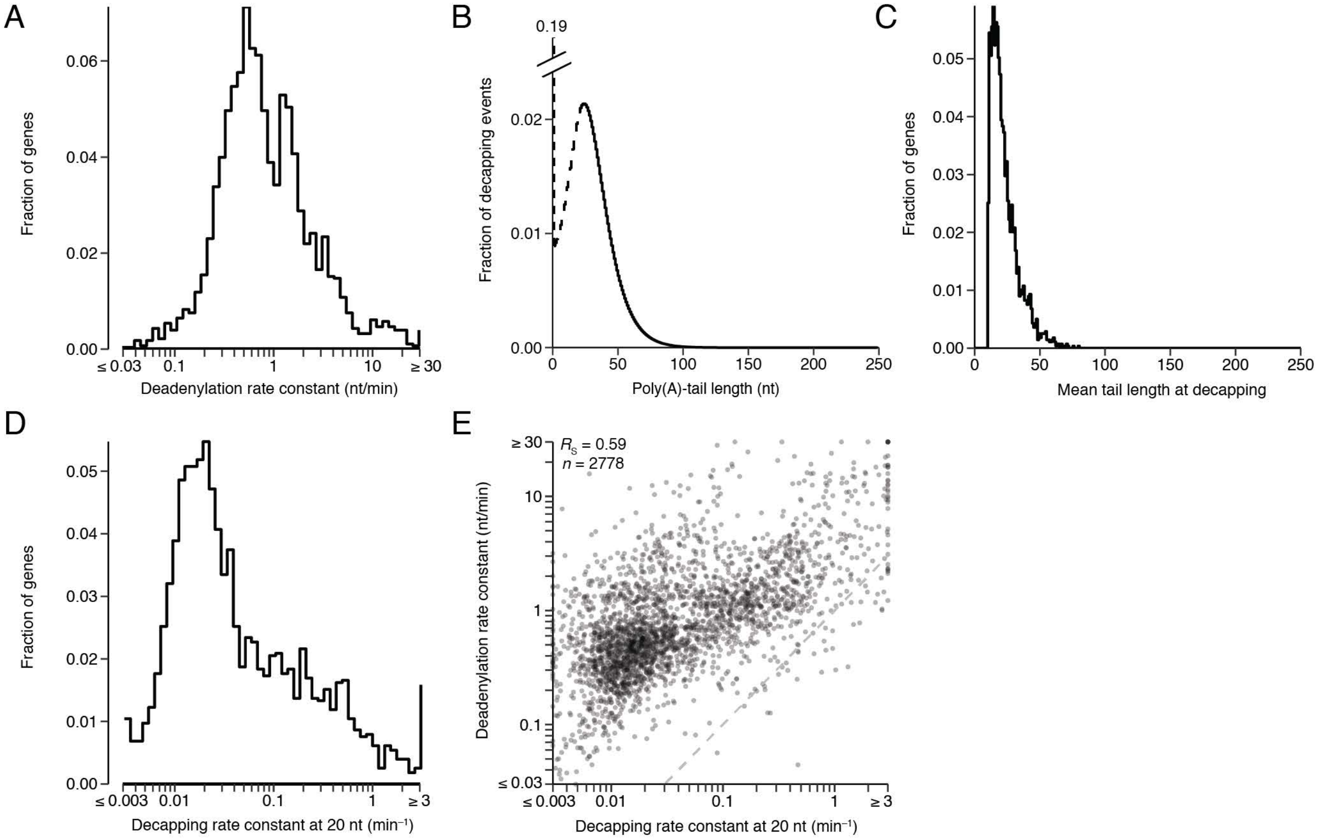

A
Fraction of genes
0.00
0.02
0.04
0.06
≤ 0.03
0.1
1
10
≥ 30
Deadenylation rate constant (nt/min)
B
0.19
Fraction of decapping events
0.00
0.01
0.02
0
50
100
150
200
250
Poly(A)-tail length (nt)
C
Fraction of genes
0.00
0.01
0.02
0.03
0.04
0.05
0
50
100
150
200
250
Mean tail length at decapping
D
Fraction of genes
0.00
0.01
0.02
0.03
0.04
0.05
≤ 0.003
0.01
0.1
1
≥ 3
Decapping rate constant at 20 nt (min⁻¹)
E
R_S = 0.59
n = 2778
Deadenylation rate constant (nt/min)
≥ 30
10
1
0.1
≤ 0.03
≤ 0.003
0.01
0.1
1
≥ 3
Decapping rate constant at 20 nt (min⁻¹)

Figure 6

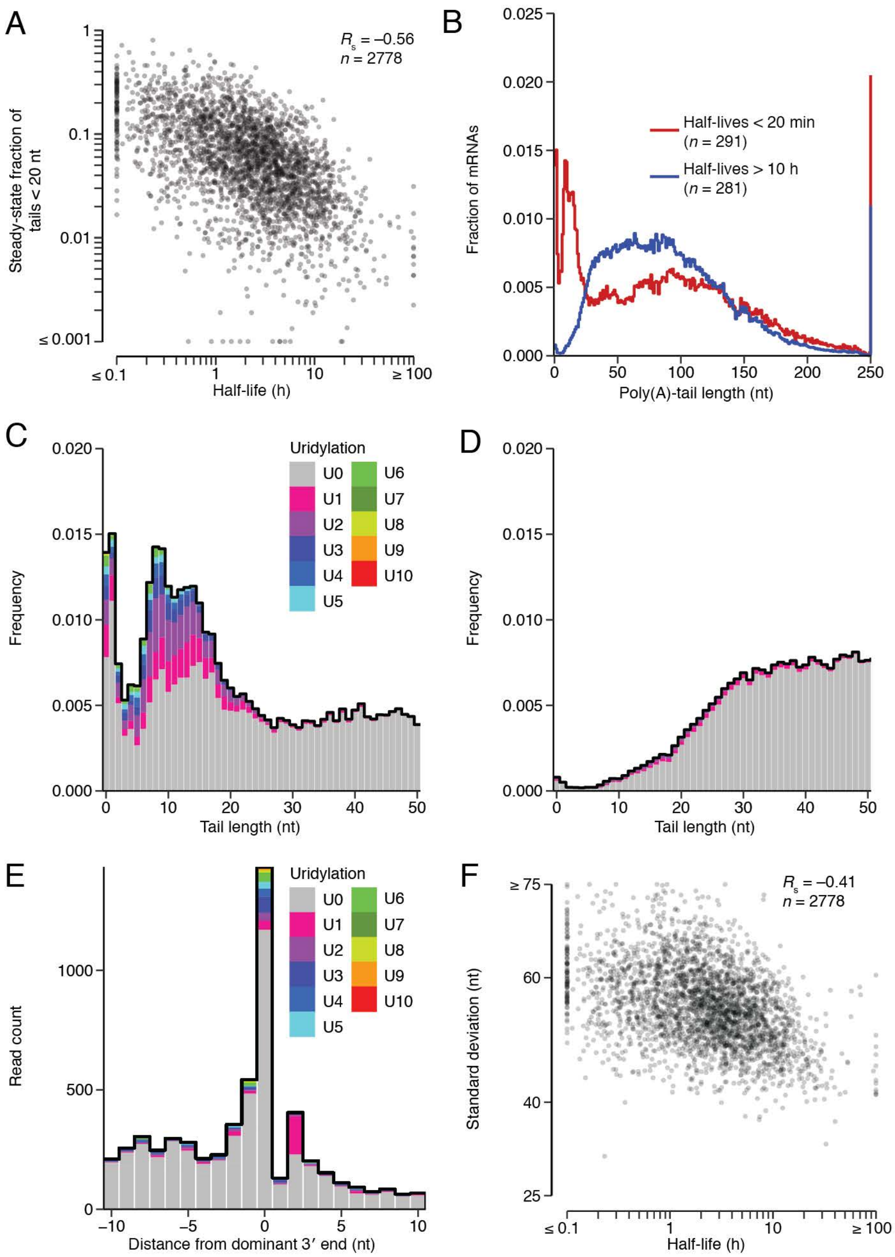

A
R_s = –0.56
n = 2778
Steady-state fraction of tails < 20 nt
1
0.1
0.01
≤ 0.001
≤ 0.1
1
10
≥ 100
Half-life (h)
B
0.025
0.020
0.015
0.010
0.005
0.000
Fraction of mRNAs
Half-lives < 20 min (n = 291)
Half-lives > 10 h (n = 281)
0
50
100
150
200
250
Poly(A)-tail length (nt)
C
0.020
0.015
0.010
0.005
0.000
Frequency
Uridylation
U0
U1
U2
U3
U4
U5
U6
U7
U8
U9
U10
0
10
20
30
40
50
Tail length (nt)
D
0.020
0.015
0.010
0.005
0.000
Frequency
0
10
20
30
40
50
Tail length (nt)
E
Read count
1000
500
0
Uridylation
U0
U1
U2
U3
U4
U5
U6
U7
U8
U9
U10
–10
–5
0
5
10
Distance from dominant 3′ end (nt)
F
R_s = –0.41
n = 2778
Standard deviation (nt)
≥ 75
60
40
25
≤ 0.1
1
10
≥ 100
Half-life (h)

Figure 7

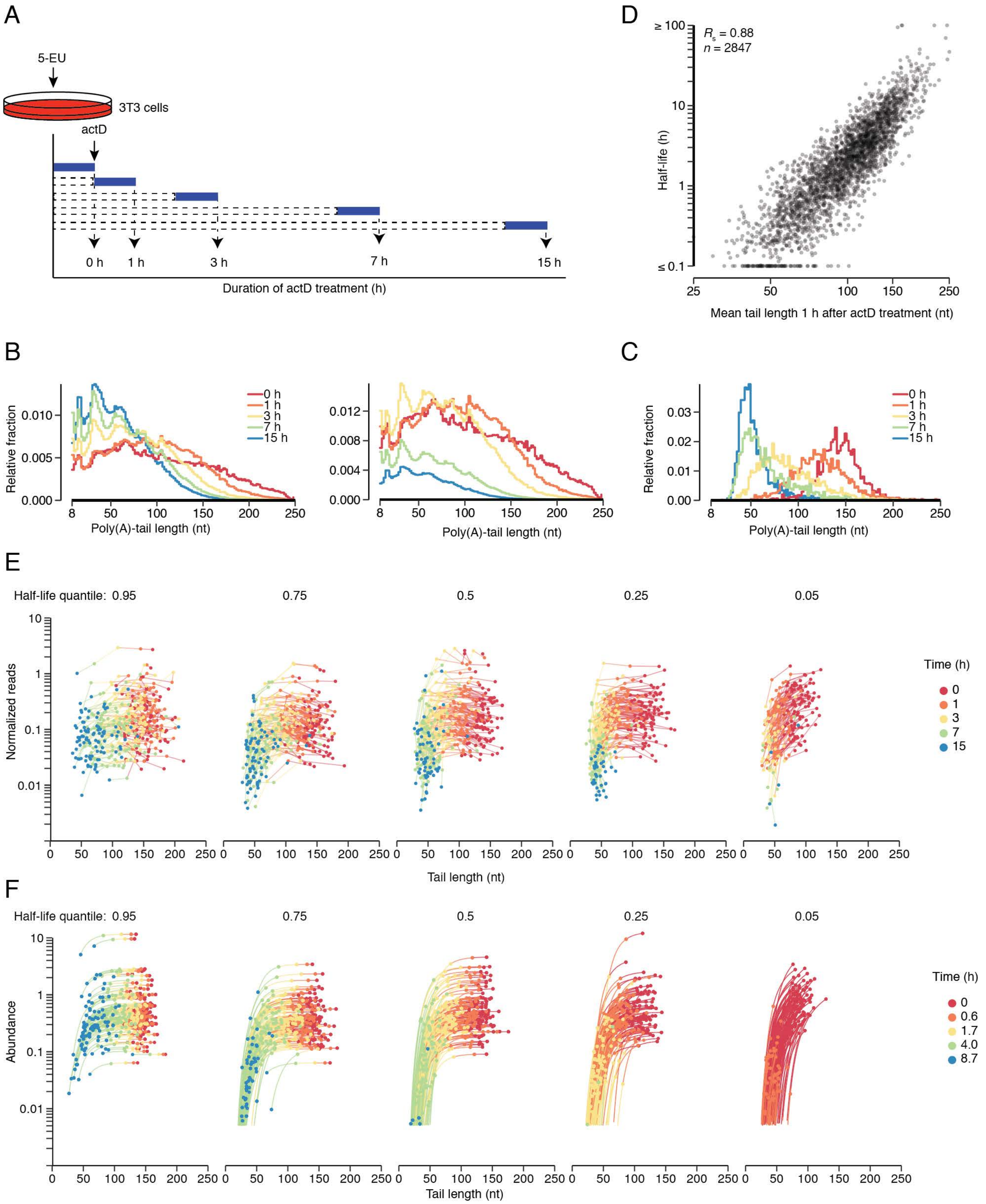

A
5-EU
3T3 cells
actD
0 h
1 h
3 h
7 h
15 h
Duration of actD treatment (h)
D
R_s = 0.88
n = 2847
≥ 100
10
1
≤ 0.1
Half-life (h)
25
50
100
150
250
Mean tail length 1 h after actD treatment (nt)
B
0 h
1 h
3 h
7 h
15 h
Relative fraction
0.010
0.005
0.000
8
50
100
150
200
250
Poly(A)-tail length (nt)
0.012
0.008
0.004
0.000
Poly(A)-tail length (nt)
C
0 h
1 h
3 h
7 h
15 h
Relative fraction
0.03
0.02
0.01
0.00
Poly(A)-tail length (nt)
E
Half-life quantile: 0.95
0.75
0.5
0.25
0.05
Normalized reads
10
1
0.1
0.01
0
50
100
150
200
250
Tail length (nt)
Time (h)
0
1
3
7
15
F
Half-life quantile: 0.95
0.75
0.5
0.25
0.05
Abundance
10
1
0.1
0.01
Tail length (nt)
Time (h)
0
0.6
1.7
4.0
8.7

Figure S1

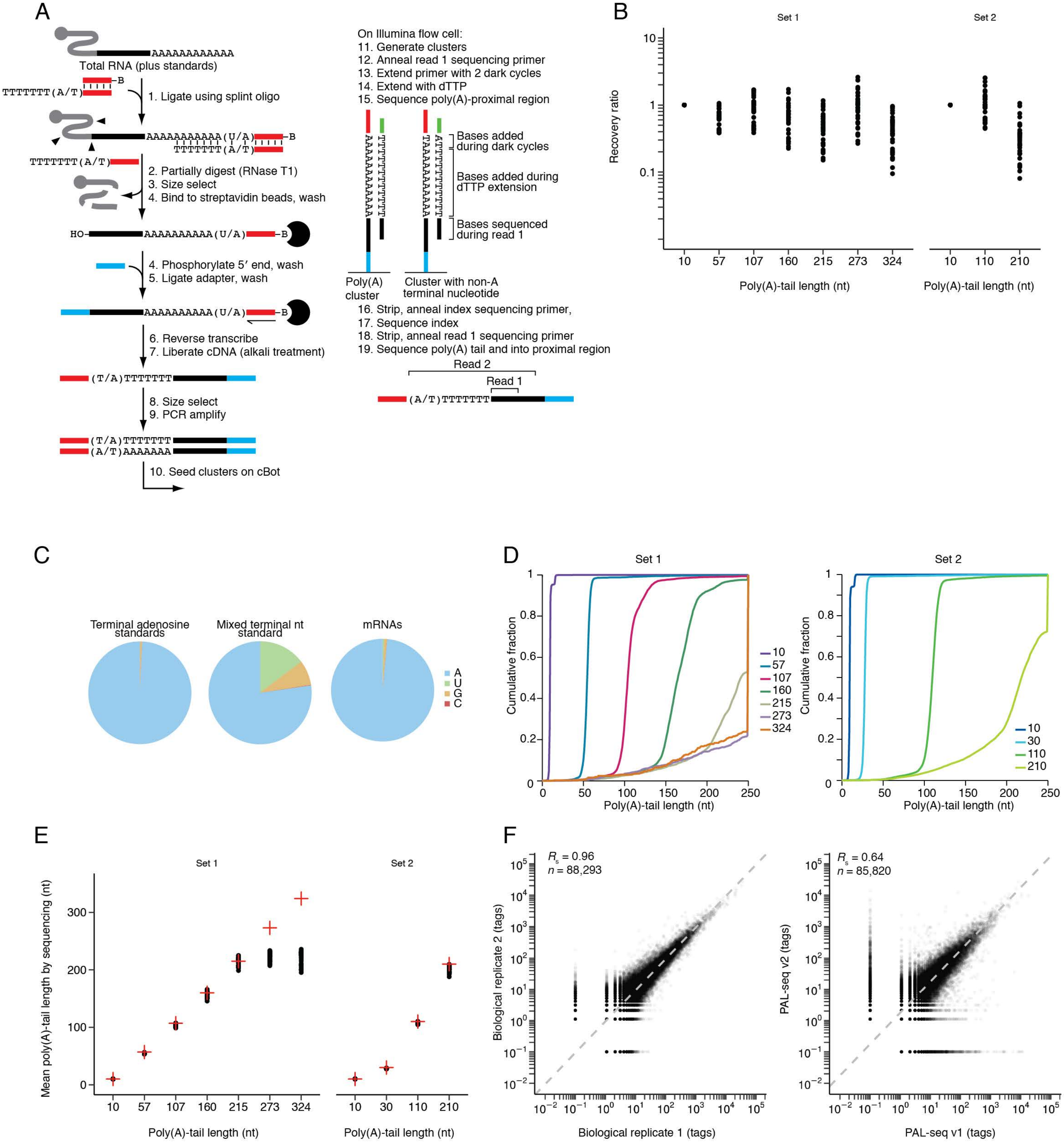

A
AAAAAAAAAAAA
Total RNA (plus standards)
TTTTTTT(A/T)
B
1. Ligate using splint oligo
AAAAAAAAAAA(U/A)
TTTTTTT(A/T)
TTTTTTT(A/T)
2. Partially digest (RNase T1)
3. Size select
4. Bind to streptavidin beads, wash
HO-
AAAAAAAAAA(U/A)
4. Phosphorylate 5′ end, wash
5. Ligate adapter, wash
AAAAAAAAAA(U/A)
6. Reverse transcribe
7. Liberate cDNA (alkali treatment)
(T/A)TTTTTTT
8. Size select
9. PCR amplify
(T/A)TTTTTTT
(A/T)AAAAAAA
10. Seed clusters on cBot
On Illumina flow cell:
11. Generate clusters
12. Anneal read 1 sequencing primer
13. Extend primer with 2 dark cycles
14. Extend with dTTP
15. Sequence poly(A)-proximal region
AAAAAAAAAAAA
TTTTTTTTTTT
TAAAAAAAAAAA
ATTTTTTTTTTT
Bases added during dark cycles
Bases added during dTTP extension
Bases sequenced during read 1
Poly(A) cluster
Cluster with non-A terminal nucleotide
16. Strip, anneal index sequencing primer,
17. Sequence index
18. Strip, anneal read 1 sequencing primer
19. Sequence poly(A) tail and into proximal region
Read 2
Read 1
(A/T)TTTTTTT
B
Set 1
Set 2
Recovery ratio
10
1
0.1
10 57 107 160 215 273 324
Poly(A)-tail length (nt)
10 110 210
Poly(A)-tail length (nt)
C
Terminal adenosine standards
Mixed terminal nt standard
mRNAs
A
U
G
C
D
Set 1
Cumulative fraction
1
0.8
0.6
0.4
0.2
0
0 50 100 150 200 250
Poly(A)-tail length (nt)
10
57
107
160
215
273
324
Set 2
Cumulative fraction
10
30
110
210
0 50 100 150 200 250
Poly(A)-tail length (nt)
E
Set 1
Set 2
Mean poly(A)-tail length by sequencing (nt)
300
200
100
0
10 57 107 160 215 273 324
Poly(A)-tail length (nt)
10 30 110 210
Poly(A)-tail length (nt)
F
R_s = 0.96
n = 88,293
Biological replicate 2 (tags)
Biological replicate 1 (tags)
10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5}
R_s = 0.64
n = 85,820
PAL-seq v2 (tags)
PAL-seq v1 (tags)

Figure S2

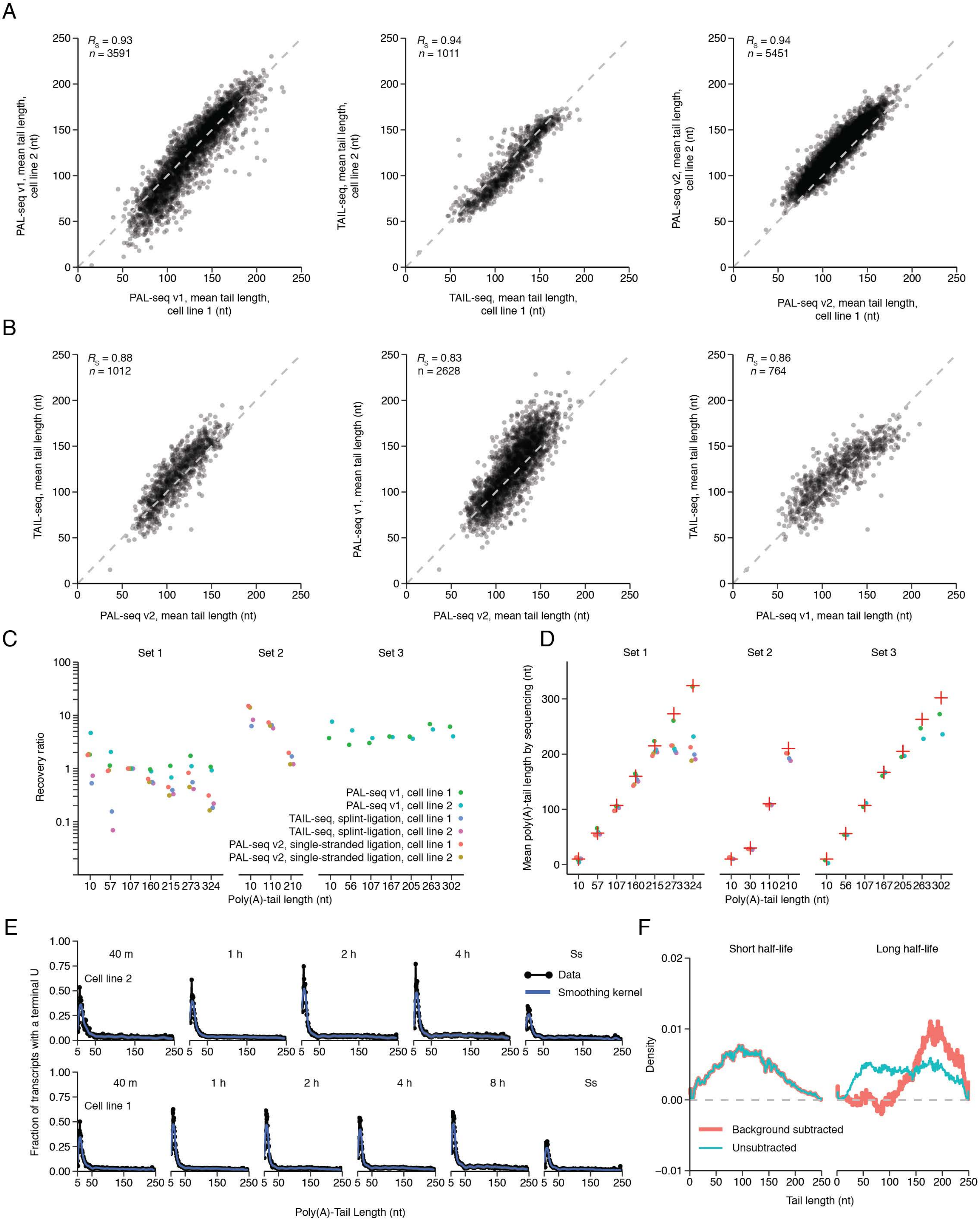

A
R_s = 0.93
n = 3591
PAL-seq v1, mean tail length, cell line 2 (nt)
PAL-seq v1, mean tail length, cell line 1 (nt)
R_s = 0.94
n = 1011
TAIL-seq, mean tail length, cell line 2 (nt)
TAIL-seq, mean tail length, cell line 1 (nt)
R_s = 0.94
n = 5451
PAL-seq v2, mean tail length, cell line 2 (nt)
PAL-seq v2, mean tail length, cell line 1 (nt)
0
50
100
150
200
250
B
R_s = 0.88
n = 1012
TAIL-seq, mean tail length (nt)
PAL-seq v2, mean tail length (nt)
R_s = 0.83
n = 2628
PAL-seq v1, mean tail length (nt)
PAL-seq v2, mean tail length (nt)
R_s = 0.86
n = 764
TAIL-seq, mean tail length (nt)
PAL-seq v1, mean tail length (nt)
C
Set 1
Set 2
Set 3
Recovery ratio
100
10
1
0.1
PAL-seq v1, cell line 1 ●
PAL-seq v1, cell line 2 ●
TAIL-seq, splint-ligation, cell line 1 ●
TAIL-seq, splint-ligation, cell line 2 ●
PAL-seq v2, single-stranded ligation, cell line 1 ●
PAL-seq v2, single-stranded ligation, cell line 2 ●
10 57 107 160 215 273 324
10 110 210
10 56 107 167 205 263 302
Poly(A)-tail length (nt)
D
Set 1
Set 2
Set 3
Mean poly(A)-tail length by sequencing (nt)
0
100
200
300
10 57 107 160 215 273 324
10 30 110 210
10 56 107 167 205 263 302
Poly(A)-tail length (nt)
E
Fraction of transcripts with a terminal U
Cell line 2
40 m
1 h
2 h
4 h
Ss
Data
Smoothing kernel
Cell line 1
40 m
1 h
2 h
4 h
8 h
Ss
0.00
0.25
0.50
0.75
1.00
5 50 150 250
Poly(A)-Tail Length (nt)
F
Short half-life
Long half-life
Density
0.02
0.01
0.00
−0.01
Background subtracted
Unsubtracted
0 50 100 150 200 250
Tail length (nt)

Figure S3

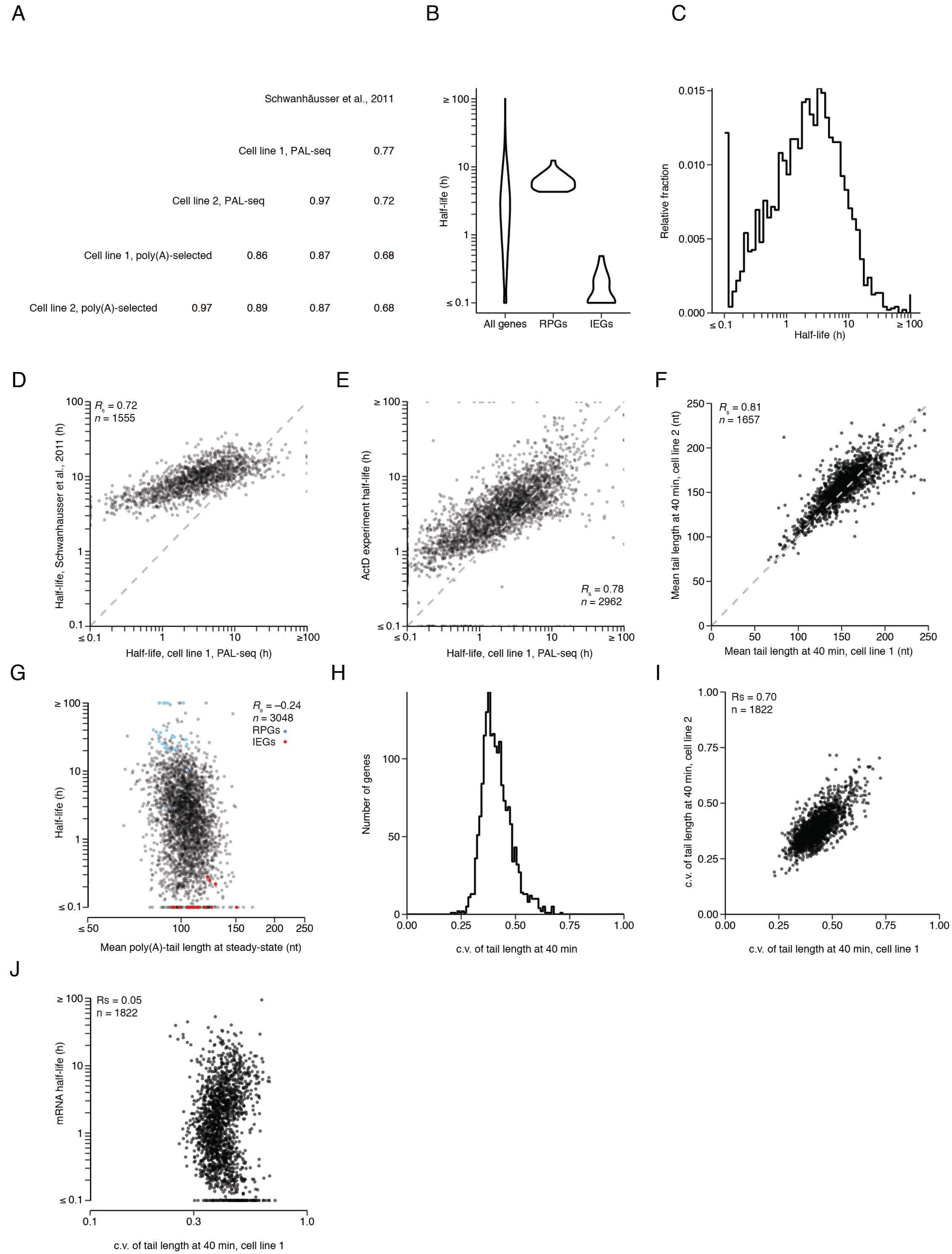

A
Schwanhäusser et al., 2011
Cell line 1, PAL-seq 0.77
Cell line 2, PAL-seq 0.97 0.72
Cell line 1, poly(A)-selected 0.86 0.87 0.68
Cell line 2, poly(A)-selected 0.97 0.89 0.87 0.68
B
Half-life (h)
≥ 100
10
1
≤ 0.1
All genes
RPGs
IEGs
C
Relative fraction
0.015
0.010
0.005
0.000
≤ 0.1
1
10
≥ 100
Half-life (h)
D
R_s = 0.72
n = 1555
Half-life, Schwanhausser et al., 2011 (h)
100
10
1
0.1
≤ 0.1
1
10
≥100
Half-life, cell line 1, PAL-seq (h)
E
ActD experiment half-life (h)
≥ 100
10
1
≤ 0.1
R_s = 0.78
n = 2962
≤ 0.1
1
10
≥ 100
Half-life, cell line 1, PAL-seq (h)
F
R_s = 0.81
n = 1657
Mean tail length at 40 min, cell line 2 (nt)
250
200
150
100
50
0
0
50
100
150
200
250
Mean tail length at 40 min, cell line 1 (nt)
G
R_s = −0.24
n = 3048
RPGs ●
IEGs ●
Half-life (h)
≥ 100
10
1
≤ 0.1
≤ 50
100
150
200
250
Mean poly(A)-tail length at steady-state (nt)
H
Number of genes
100
50
0
0.00
0.25
0.50
0.75
1.00
c.v. of tail length at 40 min
I
Rs = 0.70
n = 1822
c.v. of tail length at 40 min, cell line 2
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
c.v. of tail length at 40 min, cell line 1
J
Rs = 0.05
n = 1822
mRNA half-life (h)
≥ 100
10
1
≤ 0.1
0.1
0.3
1.0
c.v. of tail length at 40 min, cell line 1

Figure S4

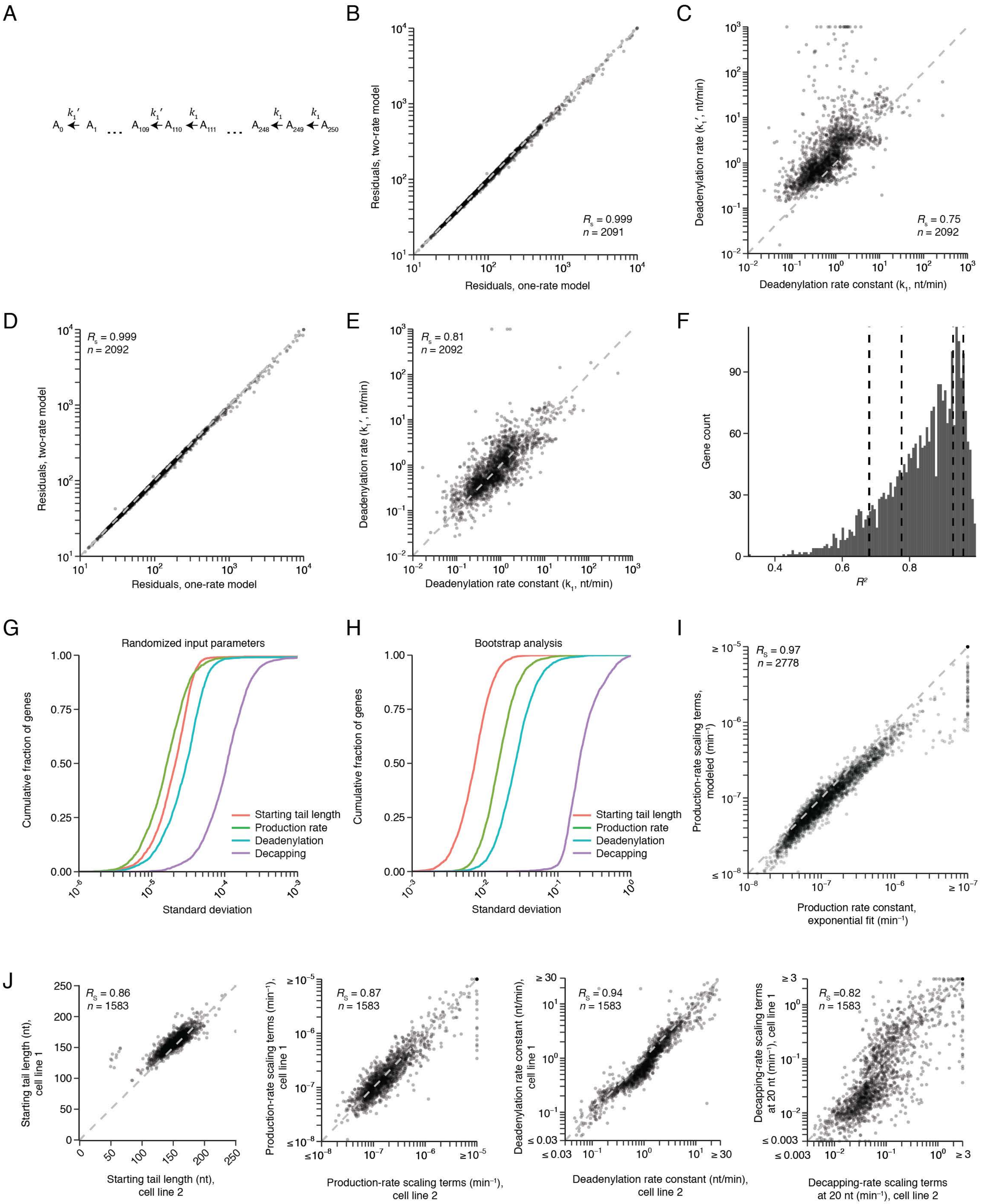
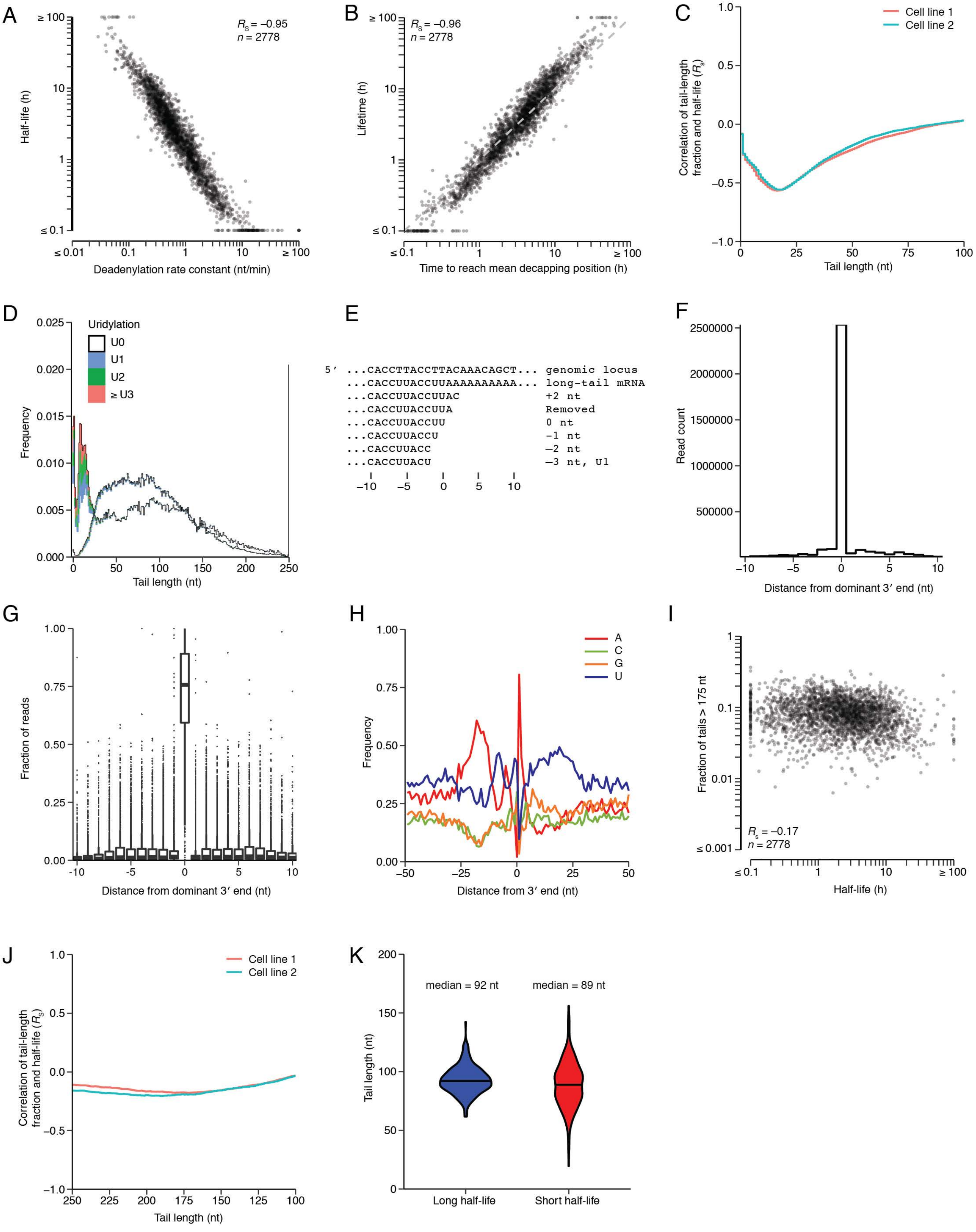

A
$A_0 \xleftarrow{k_1'} A_1 \dots A_{109} \xleftarrow{k_1'} A_{110} \xleftarrow{k_1} A_{111} \dots A_{248} \xleftarrow{k_1} A_{249} \xleftarrow{k_1} A_{250}$
B
Residuals, two-rate model
Residuals, one-rate model
R_s = 0.999
n = 2091
C
Deadenylation rate (k_1', nt/min)
Deadenylation rate constant (k_1, nt/min)
R_s = 0.75
n = 2092
D
Residuals, two-rate model
Residuals, one-rate model
R_s = 0.999
n = 2092
E
Deadenylation rate (k_1', nt/min)
Deadenylation rate constant (k_1, nt/min)
R_s = 0.81
n = 2092
F
Gene count
R^2
0 30 60 90
0.4 0.6 0.8
G
Randomized input parameters
Cumulative fraction of genes
Standard deviation
Starting tail length
Production rate
Deadenylation
Decapping
H
Bootstrap analysis
Cumulative fraction of genes
Standard deviation
Starting tail length
Production rate
Deadenylation
Decapping
I
Production-rate scaling terms, modeled (min^{-1})
Production rate constant, exponential fit (min^{-1})
R_S = 0.97
n = 2778
≥ 10^{-5}
≤ 10^{-8}
≥ 10^{-7}
J
Starting tail length (nt), cell line 1
Starting tail length (nt), cell line 2
R_S = 0.86
n = 1583
Production-rate scaling terms (min^{-1}), cell line 1
Production-rate scaling terms (min^{-1}), cell line 2
R_S = 0.87
n = 1583
Deadenylation rate constant (nt/min), cell line 1
Deadenylation rate constant (nt/min), cell line 2
R_S = 0.94
n = 1583
Decapping-rate scaling terms at 20 nt (min^{-1}), cell line 1
Decapping-rate scaling terms at 20 nt (min^{-1}), cell line 2
R_S =0.82
n = 1583

Figure S5


A
≥ 100
10
1
≤ 0.1
Half-life (h)
R_s = –0.95
n = 2778
≤ 0.01
0.1
1
10
≥ 100
Deadenylation rate constant (nt/min)
B
≥ 100
10
1
≤ 0.1
Lifetime (h)
R_s = –0.96
n = 2778
≤ 0.1
1
10
≥ 100
Time to reach mean decapping position (h)
C
1.0
0.5
0.0
–0.5
–1.0
Correlation of tail-length fraction and half-life (R_s)
Cell line 1
Cell line 2
0
25
50
75
100
Tail length (nt)
D
0.025
0.020
0.015
0.010
0.005
0.000
Frequency
Uridylation
U0
U1
U2
≥ U3
0
50
100
150
200
250
Tail length (nt)
E
5′ ...CACCTTACCTTACAAACAGCT... genomic locus
...CACCUUACCUUAAAAAAAAAAA... long-tail mRNA
...CACCUUACCUUAC +2 nt
...CACCUUACCUUA Removed
...CACCUUACCUU 0 nt
...CACCUUACCU –1 nt
...CACCUUACC –2 nt
...CACCUUACU –3 nt, U1
–10
–5
0
5
10
F
2500000
2000000
1500000
1000000
500000
Read count
–10
–5
0
5
10
Distance from dominant 3′ end (nt)
G
1.00
0.75
0.50
0.25
0.00
Fraction of reads
-10
-5
0
5
10
Distance from dominant 3′ end (nt)
H
1.00
0.75
0.50
0.25
0.00
Frequency
A
C
G
U
–50
–25
0
25
50
Distance from 3′ end (nt)
I
1
0.1
0.01
≤ 0.001
Fraction of tails > 175 nt
R_s = –0.17
n = 2778
≤ 0.1
1
10
≥ 100
Half-life (h)
J
1.0
0.5
0.0
–0.5
–1.0
Correlation of tail-length fraction and half-life (R_s)
Cell line 1
Cell line 2
250
225
200
175
150
125
100
Tail length (nt)
K
200
150
100
50
0
Tail length (nt)
median = 92 nt
median = 89 nt
Long half-life
Short half-life

Appendix 3. Effects of cooperativity on miRNA targeting

When target sites to miRNAs are sufficiently close together, it has been observed that they tend to confer more repression than would be expected by two independent sites (Doench and Sharp 2004; Grimson et al. 2007; Sætrom et al. 2007; Broderick et al. 2011). The collection of 28 RNA-seq datasets from miRNA transfection experiments with high signal-to-noise ratios (Chapter 2), allows us to probe this cooperative effect with high resolution. I started by isolating the miRNA and mRNA pairs for which the miRNA has exactly two 7mer sites (i.e. 7mer-A1 or 7mer-m8) and no other 6mer-containing sites in the mRNA 3′ UTR. I then calculated the distances between the end of the first site and the beginning of the second site for each of these miRNA–mRNA pairs. For each possible range of distances from 0 to 200, I averaged the repression values observed for miRNA–mRNA pairs for which the two sites to the miRNA in the mRNA were in that range and plotted the results (Figure 1). For example, the square at position 20 on the *y*-axis and 40 on the *x*-axis corresponds to all miRNA–mRNA pairs for which the two sites to the miRNA in the mRNA were between 20 and 40 nucleotides apart, inclusively. For mRNAs in which the two sites are closer than 15 nucleotides apart, I observed no better repression than that conferred by a single 7mer site (Figure 2), indicating that two RISC complexes cannot engage the same mRNA at the same time if they are binding sites that are 15 nucleotides apart or less. The cutoff after which two sites no longer act cooperatively is not as sharp of a boundary, presumably because mRNA secondary structure between two sites can change the effective distance between two sites. However, this upper bound appears to be somewhere between 60 and 80 nucleotides, and a strong cooperative effect was observed for sites between 15 and 60 nucleotides apart (Figure 2).

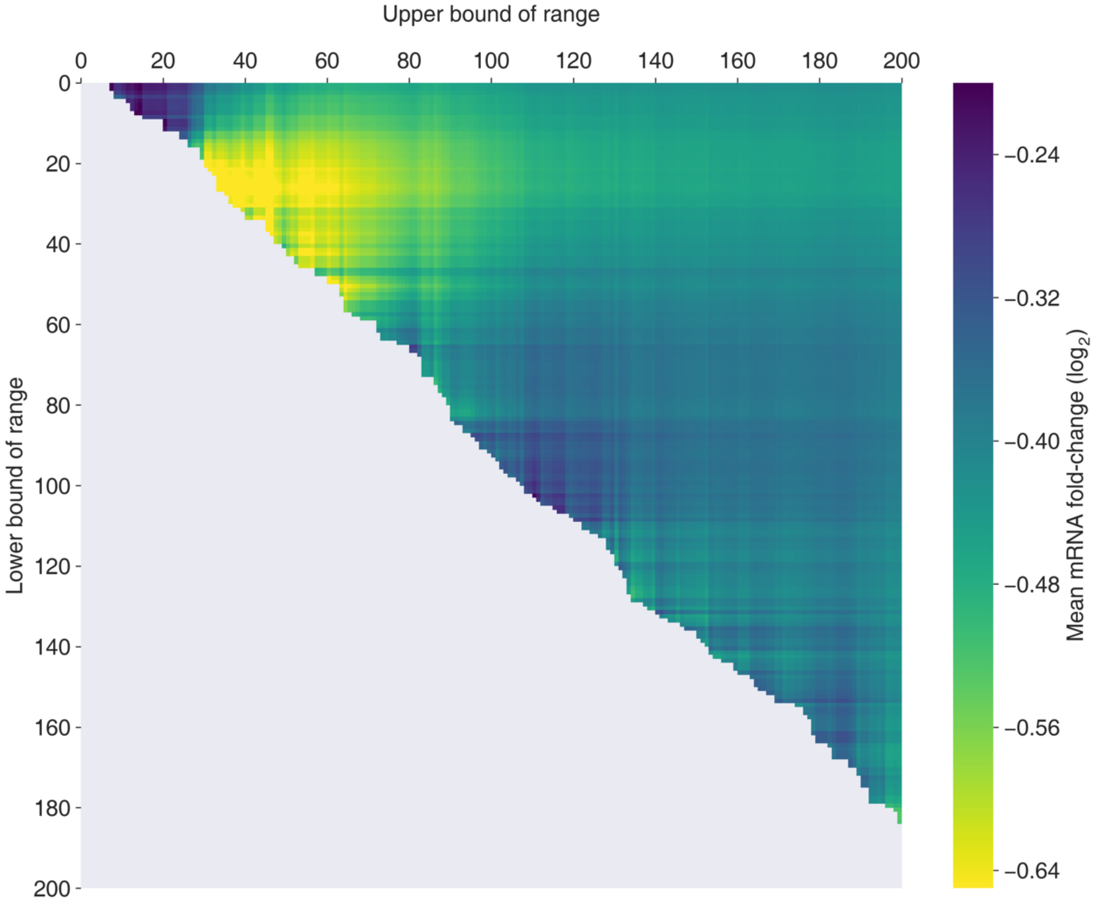


Figure 1. Exploration of mRNA fold-changes (key) conferred by two 7mer sites separated by anywhere from 0 to 200 nucleotides. Ranges with fewer than 20 examples were omitted.

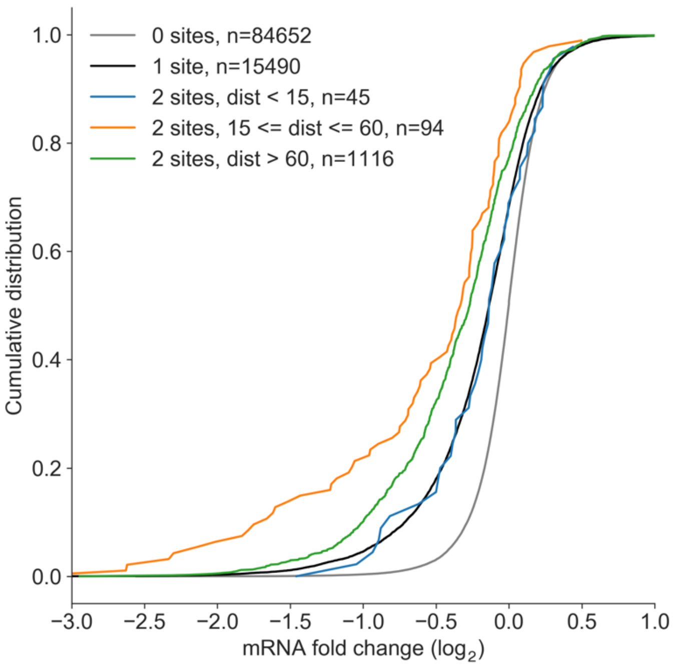


Figure 2. Cumulative distributions of mRNAs with two 7mer sites that are either too close together to both bind AGO at the same time (blue), in the optimal range for cooperativity (orange), or outside the optimal range for cooperativity (green) compared to mRNAs containing a single 7mer site (black) or no 6mer site at all (grey).

References

Broderick, Jennifer A., William E. Salomon, Sean P. Ryder, Neil Aronin, and Phillip D. Zamore. 2011. “Argonaute Protein Identity and Pairing Geometry Determine Cooperativity in Mammalian RNA Silencing.” *RNA*. https://doi.org/10.1261/rna.2778911.

Doench, John G., and Phillip A. Sharp. 2004. “Specificity of MicroRNA Target Selection in Translational Repression.” *Genes & Development* 18 (5): 504–11. https://doi.org/10.1101/gad.1184404.

Grimson, Andrew, Kyle Kai How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. 2007. “MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing.” *Molecular Cell* 27 (1): 91–105. https://doi.org/10.1016/j.molcel.2007.06.017.

Sætrom, Pål, Bret S.E. Heale, Ola Snøve, Lars Aagaard, Jessica Alluin, and John J. Rossi. 2007. “Distance Constraints between MicroRNA Target Sites Dictate Efficacy and Cooperativity.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm133.

Curriculum vitae

Education

Massachusetts Institute of Technology, 2014 – 2019
Ph.D., Computational and Systems Biology
Advisor: David Bartel

Harvard University, 2010 – 2014
AB degree Magna Cum Laude Highest Honors, Chemical and Physical Biology
Secondary Field, Computer Science

Research Experience

Ph.D. candidate, Massachusetts Institute of Technology, May 2015 – October 2019
Advisor: Professor David Bartel
Built a biochemically informed model of miRNA targeting efficacy and a convolutional neural network to predict miRNA site affinities from sequence. Techniques used include building models with TensorFlow, RNA-seq analysis, and tissue culture. Collaborated with David Page's group to develop a background model for miRNA target-site conservation.

Research intern, Computational Sciences, Pfizer, September 2018 – December 2018
Advisor: Dr. Vishnu Sresht
Adapted an existing graph-based variational autoencoder to map a latent space for small-molecule compounds against a particular target.

Research associate, Harvard University Cambridge, November 2012 – May 2014
Advisor: Professor Joanna Aizenberg
Measured and simulated the diffusion of aqueous solutions through inverse-opal nanostructures. Manufactured the nanostructures, operated a scanning electron microscope to confirm the structures, and used fluorescence recovery after photobleaching to measure diffusion of dyes through the structures.

Research associate, Harvard University Cambridge, MA
Advisor: Professor Naomi Pierce, March 2011 – August 2011
Designed and carried out an experiment to determine whether ant attendance affects larvae aggregation in Jalmenus evagoras caterpillars.

Teaching Experience

Massachusetts Institute of Technology September 2014 – December 2014
Head Teaching Assistant, Thermodynamics of Biomolecular Systems (20.110)
Taught weekly recitations and held weekly office hours to help students with course material. Helped write problem sets and grade exams.

Harvard University Bureau of Study Counsel February 2013—May 2014
Peer Tutor
Tutored fellow students in LS 1a (Introduction to molecular and cellular biology), LS 1b (Genetics), CS 50 (Introduction to Computer Science), and STAT 110 (Introduction to Probability). Tutoring was both one-on-one and in groups.

Publications

Eisen, Timothy J*, Stephen W Eichhorn*, Alexander O Subtelny*, Kathy S Lin, Sean E McGeary, Sumeet Gupta, and David P Bartel. 2019 "The Dynamics of Cytoplasmic mRNA Metabolism." *BioRxiv*, September, 763599. https://doi.org/10.1101/763599

McGeary, Sean E*, Kathy S Lin*, Charlie Y Shi, Namita Bisaria, and David P Bartel. 2018. "The Biochemical Basis of MicroRNA Targeting Efficacy." *BioRxiv*, January, 414763. https://doi.org/10.1101/414763.

Naqvi, Sahin, Daniel W. Bellott, Kathy S. Lin, and David C. Page. 2018. "Conserved MicroRNA Targeting Reveals Preexisting Gene Dosage Sensitivities That Shaped Amniote Sex Chromosome Evolution." *Genome Research*. https://doi.org/10.1101/gr.230433.117.

*These authors contributed equally to this work.