**TO DO LIST:**

**SEAN**

1. Lsy-6 doesn’t have an enhanced 6mer site
2. Systematically check the pivot-bulged sites. (Put them in supplemental Figure 2; right next to the Daeyun Beck figures.
3. Check which bulge sites are in the top list.
4. Average the Daeyun Beck sites; don’t have them all iterated out.
5. Figure 1: Hard code the Kd values.
6. Generate 16mer Kd values, by performing 12mer Kd analysis on both sites of the 8mer. Didn’t work; sticking with 12mer
7. Figure 1: Add new Figure1A, with the canonical site schematic similar to Dave’s review.
8. Add colors for all new sites.
9. Finish miR-7 site type list.
10. Resize Figure 2.
11. Supplementary Figure1: Show a figure demonstrating the drop-out R=.99. Options:
    1. Show the worst pairwise comparison.
    2. Show the 5 by 5 diagonal plot, of the correlation values.
12. Show pairwise miR-1 plot with different miR-1 preps and different random libraries.
13. Figure out cutoff for sites in Figure 1 D
14. Figure S1: Add enrichment across the randomized library
    1. miR-1 / let-7a problem?
       1. Due to the lack of I\_combined in the site counts, adding the new sites (Assigning the sites now; 5:13 PM)
    2. miR-7 problem:
       1. Also due to the lack of I\_combined sample in the site count table; (Assigning it at 5:23 PM)
    3. Get 10mers into the figure, why doesn’t it work?:
       1. Due to SiteList taking a new argument., ostensibly fixed. (5:24)
    4. Sizing
    5. Get the Kd in the legend to have a subscript.
15. Figure 2: Make a schematic for adding energy of A1 and m8.
16. Add miR-7-23nt (equilibrium2\_nb) to Figure 2
17. Add miR-7-23nt (equilibrium2\_nb) to Figure S2i (update, now S2)
18. Add miR-7-23nt (equilibrium2\_nb) to Figure S2ii (update, now S3)
19. Add miR-7 to new Figure 4 (Old figure 3), and make updates.
    1. Flanking
    2. Structure
    3. miR-1 Issue arises: The 7mer-m8w3 must be a not real site, as it has a strong preference for the context GT, on the end. This gives a ACATT**TCGT**. This occurs in the fanking sequence: **TCGT**ATG
       1. Checking this with the -1, 0, 1, and 2 n\_constant.
       2. Fit Site Kds to -3, -2, -1, 0, 1, 2, 3, 4, 5
20. Swap Figures 3 and 4.
21. Fix New Figure 3C, add trend lines to both plots, rather than connecting lines.
    1. Finish the figure before dealing with issue II:
    2. Issue: the miR-7-23nt kds changed, because the “fixed” Kd concentrations were set erroneously to the combined, global fit, rather than the global, non-combined fit.
       1. Nested Issue: when looking at the comparison of the input read counts for the combined and not combined miR-7 -23nt samples, noticed that the 0 protein for the miR-7-23nt, both equilibrium\_nb & equilibrium2\_nb is the same. Must have some mixing of the two experiments in the data frame.

Plan:

* + - 1. Check the 0 protein for both, the raw reads, look to see if they are the same. File wont’ preprocess, only the test version, seeing why.
      2. The combined input fitting to miR-7-23nt equilibrium2\_nb ostensibly fits better than the non-combined input, the combined input being not made from that input sample (\*should\* work better with noncobined ipput).  
         Trying to retain the highest concentration of miR-7-23nt (the 40% v/v sample) in the fitting, to see if this makes it work better. Also trying to see if using a different concentration of input library works better.
         1. miR-7-23nt equilibrium\_nb 0 test works.
         2. miR-7-23nt equilibrium\_nb 0 doesn’t work.

1. Decision: Should I use best canonical site, single canonical sites, or use a linear combination of all canonical sites in the UTR?
   1. Kathy: Use linear model.
   2. Dave: Fine with linear model.
2. Check the Occupancy cutoffs for all the miR-1 miR-124, miR-155, lsy-6, miR-7 sites with non-canonical sites.
   1. Became Kd and site-number cutoffs. Have a Kd cutoff of 0.1 nM, now making a site-number cutoff.
      1. Kd cutoff
      2. Site-number cutoff: Dave picked 20
   2. Add in the 6mer-m8 and 6mer-A1 into these plots
3. Figure Legends.
   1. Figure 1
   2. Figure 2
   3. Figure 3
   4. Figure 4
   5. Figure S1
   6. Figure S2
   7. Figure S3
   8. Figure S4
4. Look for seed remnants upstream of the 3p sites for all the sites: check to see what the % is of partial seeds compared to the input sites.
   1. miR-155
   2. miR-124
   3. lsy-6
5. Supplemental Math methods (Finish Tim’s comments).
6. Supplemental Experimental methods
7. Ask people which they’d rather see: Model vs Data or Data vs model: keep it consistent. Model x axis, data y axis
8. Add supplemental figure with the flanking Kds for the other four miRNAs.Delete the heatmap with the 6 canonical sites. DONE
9. Bootstrap for the linear model to show that the interaction terms are significant.
10. Make the pivot bulge sites brown-gold.
    1. Ensure that no other sites are brown-gold
11. Make sure that error bars don’t change with more or less reps. Use miR-7 to do this
    1. Figure out the combination of the leave one out cross validation, the lack of a 4% protein, and the number of reps, ans contributing to the size of the miR-7 error bar problem.
12. Change miR-1 Kds to being with the miR-1 input, rather than with the combined input.
    1. This can’t be done for miR-1 in Figure 4, because the input is too sparse. Need to use combined input for the sparsity problem, both with the accuracy of the site type Kds, and the accuracy of the average target site accessibility. Let Dave know.
13. Change miR-7 Kds to being with the miR-7 input, rather than with the combined input.
14. Ask Dave: use measured concentrations for Ago miR-1concentration, or inferred model concentrations? (Currently inconsistent in the text).
    1. Use the calculated protein concentrations, not the model concentrations, should be 7.6, 24, 76, 240, and 760 pM for AGO2–miR-1.
15. Ask Dave: Use pM label when describing the Relative Kds? They are relative, but they are literally specified as pM in the model, they just don’t mean anything. Use no units, and put in the text that it is with respect to a 1 nM reference.
16. Change Kd-site type repression figure to use U and not T.

KATHY

**INTRODUCTION**

MicroRNAs (miRNAs) are ~22-nt regulatory RNAs that derive from hairpin regions of precursor transcripts(Bartel2018). 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&Izaurralde2015). 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(Bartel2009). 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(Friedman2009). Most of these 90 broadly conserved families are required for proper development or physiology, as shown by knockout studies in mice, and many families also impact pathology in mouse models of disease or injury(Bartel2018).

A fuller understanding of these numerous biological functions requires a more accurate model of miRNA targeting, with the ultimate goal of correctly predicting the effects of each miRNA on the output of each expressed gene. Thus far, the informative predictive models have focused on mRNAs with canonical 6–8-nt sites matching the miRNA seed region(Agerwal2015) (Fig. 1A). For example, searches for conserved canonical sites can identify miRNA targets above the background of false-positive predictions(Lewis2005), with the estimated probability of conserved targeting (PCT) correlating with the efficacy of target repression(Friedman2009). However, some responsive targets are not broadly conserved, and some broadly conserved targets are more responsive than others. Thus, the most informative models have been built by training on experimental datasets that report on miRNA function, i.e., the responses of mRNAs to the addition of miRNAs. Although such models can be trained on all the features known to correlate with targeting efficacy, including the type of canonical site as well as various features of the site context, the miRNAs, and the mRNAs, even the most predictive models explain only a small fraction of the effects observed upon introducing or deleting a miRNA (*r*2 ≤ 0.15) (Agerwal2015). This low coefficient of determination observed between prediction and test data implies that either the models predict the direct effects of miRNA regulation only poorly, or that the changes observed upon perturbing a miRNA are mostly due to other causes, such as experimental noise or secondary effects of inhibiting direct targets.

In principle, miRNA target prediction might be improved with a biochemical approach that considers the dissociation constant (*K*D) of each miRNA–site interaction. However, in contrast to measurements of miRNA function, for which high-throughput methods have been routinely applied to learn the global effects of individual miRNAs on mRNA levels, protein levels, and translational efficiency(REFS), *K*D values have been determined for only a few sites of only a few miRNAs(REFS). The sparsity of the biochemical data has limited insight into how targeting might differ between different miRNAs and prevented construction of an informative biochemical model of targeting. To overcome these constraints, we have adapted RNA bind-n-seq (RBNS), a method developed for the study of RNA–protein interactions{Lambert:2014jma}, to the study of interactions between RNA and the miRNA silencing complex. With this approach and an improved computational pipeline for processing RBNS results, we acquired relative *K*D values for a vast array of binding sites of six miRNAs. Analyses of these sites and their affinities has revealed unanticipated differences between miRNAs with respect to both canonical and non-canonical targeting, as well as a striking influence of local site context. With these insights and *K*D values in hand, we developed a biochemical framework for predicting miRNA targets and targeting efficacy. This framework provided additional insights into miRNA function and predictive performance that was dramatically improved over previous methods.

**RESULTS**

**A quantitative binding-affinity profile for canonical and previously unidentified site types.**

As previously implemented, RBNS provides qualitative relative binding measurements for an RNA-binding protein and an essentially exhaustive list of binding sites{Lambert:2014jma}(REF). 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 perform AGO-RBNS, we set up five binding reactions, each with a different concentration of purified human AGO2–miR-1 (range, 7.6–760 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 9mer enrichments of two independent experiments using different AGO2–miR-1 and library preparations (fig. S1; *r*2 = 0.84).

When analyzing our RBNS results, we first examined enrichment of the canonical miR-1 sites, comparing the frequency of these sites in the RNA bound in the 7.6 pM AGO2–miR-1 sample with that of the input library. As expected from the site hierarchy observed in studies of site conservation and meta analyses of endogenous site efficacy (Bartel2009), the 8mer site (perfect match to miR-1 nucleotides 2–8 followed by an A) was most enriched (39.9 fold), followed by the 7mer-m8 site (perfect match to miR-1 nucleotides 2–8, enrichment 15.5 fold), then the 7mer-A1 site (perfect match to miR-1 nucleotides 2–7 followed by an A, enrichment 7.82 fold), and the 6mer site (perfect match to miR-1 nucleotides 2–7, enrichment 3.52 fold) (Fig. 1A and C). Virtually no enrichment was observed for either the 6mer-A1 site (perfect match to miR-1 nucleotides 2–6 followed by an A, enrichment 1.02 fold) or the 6mer-m8 site (perfect match to miR-1 nucleotides 3–8, enrichment 0.97 fold) (Fig. 1A and C), again consistent with their weak signal in previous analyses of conservation and efficacy {Friedman:2008km}{Agarwal:2015bw}{Kim:2016bo}.

Analysis of enrichment of these six canonical sites across all five AGO2–miR-1 concentrations illustrated two hallmarks of this experimental platform{Lambert:2014jm}. First, as the concentration increased from 7.6 pM to 76 pM, enrichment for each of the six site types increased (Fig. 1D), which was attributable to an increase in specific 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 76 pM, which was the point at which the AGO2–miR-1 concentration matched that of the 8mer-containing molecules, 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 site types{Lambert:2014jm}. 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, with the lower AGO2–miR-1 concentrations providing greater discrimination between the high-affinity site types and the higher AGO2–miR-1 concentrations providing greater discrimination between the low-affinity site types, and no concentration providing results that quantitatively reflected differences in relative binding affinities.

The established method for inferring relative *K*Dvalues from RBNS results uses data from only one protein concentration{Lambert:2014jm} and thus cannot account for either background binding or ligand saturation. Therefore, we developed a computational strategy that simultaneously incorporates information from all concentrations of an RBNS experiment to calculate relative *K*D values. Underlying this strategy is a model that takes as input the stock concentration of purified AGO2–miR-1, the *K*D values for each miRNA site type (including the “no-site” type), and a constant amount background library recovered in all samples. With these inputs, the model outputs the predicted enrichment of each site type across the concentration series. Using this model, we performed maximum likelihood estimation (MLE) to fit the *K*D values. With these fitted *K*D values, the model explained the observed data well (Fig. 1D, lines and points, respectively). Moreover, these *K*D values were robustly estimated, as indicated by comparing *K*D values obtained using results from only four of the five AGO2–miR-1 concentrations (*r2* ≥ 0.992 for each of the ten pairwise comparisons fig. S1B). Although the quantitative binding affinities followed the same hierarchy as observed for site enrichment, the differences in affinities were of greater magnitude (Fig. 1C). For example, the binding affinity of library molecules with an 8mer site was 560-fold greater than that of molecules without a site, whereas the ratio of the enrichment of the 8mer site to that of molecules without a site ranged from 41–153-fold, depending on the AGO2–miR-1 concentration.

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(Bartel2009). However, the ability to calculate the *K*D of any *k*-mer of length ≤12 nt provided the opportunity for a *de novo* search for sites, without bias from any previous knowledge, and indeed, without even considering the miRNA sequence. In this search, we 1) calculated the enrichment of all 10mers in the bound RNA, using results from the reaction with 72 pM AGO2–miR-1, 2) determined the *k*-mer length for which the top-most enriched *k*-mer exhibited the greatest relative enrichment compared to the second-most enriched *k*-mer, 3) designated the top-most enriched *k*-mer at this length as the binding site, and 4) removed all reads containing this newly identified site from both the bound and input libraries. These four steps were iterated until no 9-nt motif remained that was enriched ≥ 2-fold. We then fit the binding model to this expanded list of site-types (Fig. 1E). This unbiased approach demonstrated that the 8mer, 7mer-m8, and 7mer-A1, and 6mer sites to miR-1 were indeed the highest-affinity site types of lengths ≤11 nt, and also identified novel site types with binding affinity resembling that 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 or a mismatched U at position 6, a bulged U somewhere between positions 4 and 6, or a mismatch at position 5 and achieve affinity ranging from 5–10 fold above the remaining no-site reads. We also observed that the motif “ACACACA” conferred binding that was 6-fold above background, which was surprising because its contiguous complementarity to miR-1 did not extend beyond two UGU segments at positions 6–8 and at 18–20. Nonetheless, of the 16,384 possible 7-nt motifs, this was the only one with >2-fold enrichment that 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 type (Fig. 1G). The 8mer site occupied 3.5–17% 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 throughout the concentration course. In aggregate, the marginal site types, including the 6mer-A1, 6mer-m8, and X noncanonical site types, occupied X–XX% of the AGO2–miR-1 complex. Moreover, because of their very high abundance, library molecules with no identified site occupied 35–60% 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:2014cz}{Denzler:2016eh}.

Together, our results confirmed that AGO2–miR-1 binds the 8mer, 7mer-m8, 7mer-A1, and 6mer site types most effectively and revealed the relative binding affinities and occupancies of these sites. In addition our results uncovered some weak yet specific affinity to another X sites that in aggregate occupied 7–10 percent of the silencing complex. Although alternative binding sites for miRNAs have been proposed based on high-throughput in vivo crosslinking studies {Loeb:2012bc}{Helwak:2013ga}, 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 non-canonical 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(REFS). We also analyzed the *lsy-6* miRNA of nematodes, which is thought to bind unusually weakly to its canonical sites(REF). As for miR-1, for each of these miRNAs we purified the AGO2–miRNA complex, performed AGO-RBNS, identified sites enriched in the bound molecules, and determined the binding affinities for each of these sites, yielding five additional sets of site-affinity profiles (Fig. 2A–D).

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 all but one of the non-canonical sites (Fig. 2A). As with miR-1, the non-canonical 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 were at positions different than those observed for miR-1, with different mismatched-nucleotide identities. The let-7a, analysis also identified two sites that, as with the miR-1 ACACACA site, could not be explained by pairing to the miRNA. These rare sites that lacked substantial pairing to the miRNA always differed for different miRNAs, which ruled out binding to a common contaminant in our AGO2–miRNA preparations.

The site-affinity profiles of miR-124, miR-155, lsy-6, and miR-7 resembled those of miR-1 and let-7 in some respects. As expected, they all included the six canonical sites. They also included noncanoncial sites with extensive yet imperfect pairing to the miRNA seeds, and again these imperfections tended to occur at different positions for different miRNAs, with different mismatched-nucleotide identities (Fig. 2B–D). However, in contrast to the miR-1 and let-7a non-canonical sites, more of the non-canonical sites of miR-155, miR-124, and lsy-6 had affinities intermingled with those of the top four canonical sites. Moreover, the profiles for these three miRNAs also included sites with extended (9–11-nt) perfect complementarity to the miRNA 3′ region, indicating that these three miRNAs have an alternative binding mode dominated by extensive pairing to the 3′ region rather than to the seed (Fig. 2B–D). These could not be classified as 3′-supplementary or 3′-compensatory sites, as they were associated with little more than chance complementarity to the miRNA seed. Therefore, we named them 3′-only sites.

In some respects the 3′-only sites resembled a type of noncanonical sites known as centered sites, which are reported to function in mammalian cells (REF). 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 (REF). Our unbiased search for sites did not identify centered sites for any of the six miRNAs. To perform a more directed search, and to investigate the region of each miRNA to which extensive non-canonical pairing is most favored, we reanalyzed the results for each miRNA, 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 reference, we also re-plotted the relative *K*D values for sequences with canonical 8mer, 6mer, and 6mer-m8 sites. For miR-155, miR-124, and lys-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, none of the 11-nt sites conferred stronger binding than did the 6mer, and for miR-7, none conferred stronger binding than did the 6mer-m8. Moreover, for all six miRNAs, the 11-nt sites that started at positions 3 or 4 and extended through the center of the miRNA conferred binding no more than 2.1-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 these sites function for some miRNAs and not others. Indeed, the newly identified 3′-only sites functioned for only miR-155, miR-124, and lys-6, and even among these, the optimal region of pairing differed, with miR-155 preferring pairing to positions 13–23, rather than 9–19 (Fig. 3A).

We next considered an expanded set of site types identified through a meta analysis of repression data{Kim:2016bo}. Supplementing 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) are proposed to substantially extend the scope of miRNA–mRNA regulatory interactions{Kim:2016bo}. We found no general support for the offset 7mer site, as the relative binding conferred by this site over the canonical 6mer-m8 nested within it ranged between 0.52 and 1.61-fold (fig. S2A–F). Among the four CDNST site types, CDNST 1 (equivalent to the 5mer-m2.6) was selected in our *de novo* identification of let-7a and miR-1 sites, where it conferred 8.7- and 3.X-fold greater affinity over no site–containing reads, respectively (Fig. 1G). However, for other miRNAs it was associated with less activity, as were the other CDNST–miRNA combinations (fig. S2A–F).

Another type of non-canonical site proposed to mediate widespread targeting is the pivot–bulge site{Chi:2012jm}. This site has canonical pairing to the seed region, except 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:2012jm} (fig S3A). Our de novo search for sites provided strong support for pivot–bulge sites of miR-124 and lys-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, fig S2iiA) is a 9-nt pivot–bulge site with affinity exceeding that of the canonical 7mer-A1 site, and the lys-6 8mer-bA(6.7) is 9-nt pivot–bulge site with affinity matching that of the canonical 7mer-m8 site (Fig. 2C–D). However, even though these pivot–bulge sites for miR-124 and lsy-6 were among the highest-affinity noncanonical sites identified, we did not identify pivot–bulge sites for any of the other three miRNAs (Fig. 1F, Fig. 2A–B), and a systematic analysis of all possible single-nucleotide bulges at each position confirmed that the pivot–bulged sites to miR-1, let-7a, and miR-1 conferred no better binding than the canonical 6mer-A1 site nested within them (fig 3B). Thus, our results supported the pivot–bulge sites proposed for two of the six miRNAs but called into question the generality of this non-canonical site type.

In addition to the differences in non-canonical 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 > 11-fold lower than that of the 7mer-m8 and approached that of the 6mer-m8 site. 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:2012df}{Salomon:2015fb}, our results show that the identity 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 five miRNAs in hand, we examined the relationship between the A at target position 1 (A1) and the match at miRNA position 8 (m8), which flank pairing to the core seed. We found that the apparent binding energy of the m8 sequence feature is consistent when measured using either the *K*D,rel between the 6mer and the 7mer-m8, or between the 7mer-A1 and the 8mer, and as well for the A1 sequence feature when measuring the *K*D,rel between the 7mer-A1 and 7mer, or between the 8mer and 7mer-m8 (Fig. 3B). These results indicated that the A1 and m8 are independent, and therefore that the relationship between the 6mer, 7mer-A1, 7mer-m8, and 8mer constitute a linear, thermodynamic cycle.

Analyses of the effects of miRNAs transfected into cells show that miRNAs with greater predicted seed-pairing stability tend to be more effective at target repression (REF). However, the relationship between the predicted pairing stability and site affinity had not been investigated. To examine this relationship, we calculated the predicted free energy of site pairing (REF) and examined the correspondence between site *K*D values and these predicted ∆*G* values (Fig. 3C). We focused on the 6mer and 7mer-m8 sites, because they lack the A1, which does not pair to the miRNA (Fig 1A). As expected, for both site types affinity increased with increased predicted pairing stability (*p =* 0.XXX and 0.XXX, respectively). However, the rate of these increases were significantly less than expected from thermodynamic principles stipulated by ∆*K*D = e–∆∆G/RT (*p =* 0.XXX and 0.XXX, respectively), which indicated that in remodeling the thermodynamic properties of the loaded miRNAs, AGO helps to even out intrinsic differences in seed-pairing stabilities that would otherwise impose much greater inequities between the targeting efficacies of different miRNAs. Thus, although lsy-6, which is known for its poor predicted seed-pairing stability(REF), did indeed have the weakest site-binding affinities of the six miRNAs, the difference between its binding affinity and that of the other miRNAs was less than might have been expected.

**Correspondence between affinity measured by AGO-RBNS and repression observed in the cell.**

To validate our in vitro binding results and evaluate their relevance for miRNA-mediated repression *in vivo*, we examined the relationship between our *K*D measurements and the repression of endogenous mRNAs after miRNA transfection into HeLa cells. When examining repression of mRNAs containing a single 6–8-nt canonical site to the transfected miRNA within their 3′ UTRs, we observed a striking relationship between AGO-RBNS–determined *K*D values and mRNA fold changes (Fig. 3D–I, *r*2 = 0.92–0.99). 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. 3X,Y). 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. These included the pivot–bulge sites for miR-124 and lsy-6, and the XXXX site for miR-X. The consistent relationship between in vitro binding affinity and intracellular repression supports a model in which repression is a function of site occupancy and 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 (REF), but the biochemical basis of this phenomenon had not been investigated, presumably because of the sparsity of affinity measurements. To overcome this limitation, we separated the 8mer site into 256 different 12-nt sites, based on the dinucleotide sequences immediately flanking each side of the 8mer (e.g., AA-8mer-AA, AA-8mer-AC, …, UU-8mer-UU), and determined *K*D values for each of the 256 sites (Fig. 4A). This analysis revealed a 100-fold range in *K*D values, depending on the identities of the flanking dinucleotides of the 8mer (0.44 and 48.3 pM for AU-8mer-UA and GG-8mer-GG, respectively), 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 miRNAs (fig. S4A–E) yielded similar results. Because the effect of flanking-dinucleotide context often exceeded the affinity differences observed between miRNA-site types, the affinity of weaker sites in more favorable contexts often exceeded that of stronger sites in less favorable contexts (Fig. 4B, fig. S4A–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 contribution of each of the four nucleotides of the two flanking dinucleotides. The output of the model agreed well with the observed *K*D values (Fig. 4C left, *r*2 = 0.94), which indicated that the effects of the flanking dinucleotides were largely consistent between miRNAs and between site types of each miRNA. A and U nucleotides each enhanced binding 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 approach the central RNA-binding channel of AGO, 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 reflected the propensity of these nucleotides to stabilize RNA secondary structure that occludes binding of the silencing complex. To investigate a 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 site accessibility previously optimized on data examining miRNA-mediated repression (REF). This score is based on the predicted probability that the target segment across from miRNA nucleotides 1–15 is unpaired. We found that predicted accessibilities of sites in the bound libraries were substantially greater than that for sites in the input library and that the difference was greatest for the samples with the lower AGO2–miR-1 concentrations (fig. S4F), as expected if the accessibility score was predictive of site accessibility and if the most accessible sites were the most preferentially bound.

Encouraged by 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, and then the sites were differentiated based on their flanking dinucleotides into 256 12-nt sites, and the geometric mean of the site-accessibility scores of each of these extended sites was compared with the RBNS-derived *K*D value (Fig. 4D). A striking correlation was observed (*r*2 = 0.80, *p* < 10−15), with all 16 sites containing a 5′-flanking GG dinucleotide having both unusually poor affinities and unusually low accessibility scores. Optimization of the accessibility score showed that predicting the pairing probability of 10-nt window opposite miRNA positions 1–10 achieved a slightly greater correlation with *K*D (fig. S4G, *r2* = 0.83).

The high correspondence of these two orthogonal values—one being the averaged result of a computational algorithm applied to reads from the input library, the other being a biochemical constant derived from RBNS analyses—strongly implied that site accessibility is the primary cause of the different binding affinities associated with flanking-dinucleotide context. However, these results did not rule out the formal possibility that the AGO–miRNA complex directly binds the flanking nucleotides with a preference that happens to correlate with predicted site accessibility. To investigate this possibility, we examined how well a primary-sequence preference would recapitulate the predicted site accessibility. Input reads that contained a miR-1 8mer site were sampled to match the frequency of each of the 256 extended sites in the 7.6 pM AGO2–miR-1 library, and the distribution of site-accessibility scores was examined. This distribution was found to resemble that of the input library much more than that of the bound (Fig. 4E, left), which disfavored the directed-binding explanation. Similar results were observed when extending the analysis to the other four AGO2–miR-1 samples (Fig. 4E, right). By contrast, sampling 8mer-containing reads from the input to match the site-accessibility scores of the bound reads nicely reproduced the flanking dinucleotide frequencies (*r*2 = 0.79) (Fig. S3H). Taken together, these results demonstrate that local sequence context has large influence on miRNA–target binding affinity, and that this influence predominantly results from the differential propensities of flanking sequences to form structures that occlude site accessibility.

**A biochemical model highly predictive of miRNA targeting**

The correspondence between our affinity measurements obtained in vitro and the repression observed in cells suggested that our measurements might provide a new resource for quantitatively predicting the effects of miRNAs in cells. In principle, miRNA targeting efficacy should be a function of the biochemical occupancy of the miRNA on its target sites. Indeed, some early target-prediction algorithms use a biochemical framework to score miRNA target predictions based on predicted occupancy of the silencing complex (Krek et al., 2005). However, approaches in which computational models are fit to expression changes observed upon introducing miRNAs have yielded more accurate ranking of targeting efficacy (Bartel2009;Agarwal2015). We reasoned that these more accurate rankings might be attributable to the ability to train on experimental data rather than having to rely on nearest-neighbor or similar free-energy models to approximate miRNA–target affinities. With this in mind, we returned to the biochemical framework to examine its performance, now that we had an exhaustive set of measured *K*D values for six miRNAs as well as RNA-seq data examining the response of cellular mRNAs after introducing each of these six miRNAs into HeLa cells.

As input for the model, we calculated for each miRNA the *K*D values for all 212,912 12-nt sequences with at least four nucleotides of contiguous pairing to the extended seed (Fig. 5A). These 212,912 sequences included many versions of each of the canonical sites, each flanked by different dinucleotides (or tri- or tetranucleotides, depending on the size and position of the site), as well as many versions of most of the non-canonical sites that we had identified (Fig. 1F, Fig. 2). We then mapped these 12-nt sequences to mRNA 3′ UTRs and applied a statistical mechanics framework (Fig. 5B, equation 1) to calculate the expected number of AGO–miRNA complexes bound per mRNA molecule at steady-state. This calculation of miRNA occupancy required an estimate of the concentration of loaded, unbound AGO in the cell ([*A*], equation 1), which was acquired by fitting a global value for each transfection experiment that best explained the observed mRNA changes. Because let-7a was already highly expressed in HeLa cells, both an endogenous [*A*] and a transfected [*A*] ([*A*e] and [*A*t], respectively) were fit for the let-7a experiment. We then used the miRNA-occupancy value and a first-order equilibrium model to calculate the expected fold-change of each mRNA (Figure 5B, equation 2). This calculation required an estimate of how much a single bound RISC complex affects the mRNA decay rate (*b*, equation 2), which was fit as a global value for all six datasets. In total, this biochemical framework takes as input the experimentally determined *K*D values, one global fitted value for *b*, and one additional fitted parameter [*A*] for each transfection dataset (two in the case of let-7a).

In light of our observations regarding the function of noncanonical sites, we wanted this model to be site-type agnostic. Whereas previously the expression of an mRNA without the influence of the transfected mRNA was estimated as its average expression in datasets in which it lacked a canonical site to the transfected miRNA (Fig. 3), this value was instead estimated as the intercept of the fit the relative change in our predictor against the relative change in the transcript expression (fig. SX). This approach had the added benefit of allowing inclusion of transcripts that had canonical sites to all six of our miRNAs.

A striking correspondence was observed when comparing mRNA changes measured upon miR-1 transfection to those predicted by the model (Fig. 5C, *r*2 = 0.54). A similar performance was observed when applying the model to *K*D values for each of the other five miRNA (fig. S5A, *r*2 = 0.43–0.61) and when plotting the combined results for all six miRNAs (Fig. 5D *r*2 = 0.53). When provided *K*D values for only the XXXXX 12-nt sequences that contained the six canonical sites, the model captured somewhat less variance (Fig. 5E *r*2 = 0.XX), which further confirmed that the noncanoncial sites mediated a minor but perceptible fraction of intracellular repression. As expected, the model performed much worse when using *K*D values for one miRNA to predict repression of another miRNA (Fig. 5F, *r*2 = 0.19). The observation of this non-zero *r*2 for the shuffled predictions presumably reflected the fact that longer UTRs are likely to contain more sites for any given miRNA, creating a moderate correlation of repression of the same gene by different miRNAs. Indeed, long UTR length alone predicted repression with an *r*2 of 0.23.

The performance of our biochemical model substantially exceeded that of the 30 target-prediction algorithms previously tested on changes in levels of endogenous mRNAs observed in response to miRNA transfection(Agarwal 2015). Part of this difference might be attributed to the improved datasets used to evaluate our model. Compared to microarray datasets used to train and test previous target-predication algorithms, our new RNA-seq datasets had stronger signal over background, in part because of more efficient transfection. To evaluate the possibility that our improved datasets explained our improved performance, we generated 11 additional datasets and iteratively retrained the latest TargetScan model on the collection of 17 miRNA-transfection datasets, leaving out one dataset each time in a 17-fold cross-validation. As with the biochemical model, the predictor was fit to the relative change in expression across transfections. Training and testing the TargetScan model on the improved datasets did increase its performance, yielding an *r*2 of 0.22–0.23 (Fig. 5G, fig. S5c), which was above that originally reported (*r*2 = 0.14) (Agarwal, Bell, Nam, & Bartel, 2015) but still substantially less than the *r*2 of 0.53 achieved by the biochemical model trained on measured binding affinities (Fig. 5D).

Although the miRNA target prediction is sometimes considered a binary classification problem, experimental results (e.g., Fig 5C) show that designation of mRNAs as either “targets” or “non-targets” is not possible. With this graded response of the transcriptome, the *r*2 values provide an appropriate metric of prediction performance but still might not accurately reflect the ability of the algorithm to predict the most responsive targets, which are typically of greatest relevance to biologists seeking insight from target predictions. Moreover, the *r*2 metric is less informative for more physiological datasets, such as those obtained after knocking out an endogenous miRNA, in which the signal from the miRNA can be dwarfed by other variability, such as that arising from experimental noise or the secondary effects of losing the miRNA. With these considerations in mind, a metric more suitable for comparing the performance of miRNA-target predictions has been developed, in which the average repression of the top-ranked predictions is compared across a broad range of thresholds(Khorshid et al, 2013). Using this metric, we found that performance of the biochemical model exceeded that of TargetScan when evaluated using the six transfection datasets (Fig. 5H) and matched that of TargetScan when evaluated using microarray data reporting de-repression of miR-155 targets in miR-155 knock-out Th1 cells in mouse (Rodriguez et al., 2007) (Fig. 5I). We used the fitted *b* value from our model, but we used a lower log([*A*]) of -7 (which is similar to the initial let-7a concentration value) as a rough estimate of the unbound concentration of AGO:miR-155 in mouse Th1 cells. Although re-repression by other endogenous miRNAs is expected to dampen the de-repression signal, particularly for transcripts with long UTRs, we still observe an improvement of our binding affinity-based predictions over TargetScan7 predictions (Figure 5K).

**Convolutional neural network for predicting site *K*D values from sequence**

Our findings that binding preferences differ substantially between miRNAs and that AGO2 seems to modulate these differences in ways that existing RNA duplex free-energy models cannot predict posed a major challenge for applying our biochemical framework to other miRNAs. Because performing RBNS for each of the known miRNAs would be impractical, we attempted to predict miRNA–target affinity from sequence using the RBNS and transfection datasets already collected. The combinatorial nature of this problem made neural networks a natural choice for this task.

The overall model had two components. The first was a convolutional neural network that predicted *K*D values from miRNA and 12-nt sequences, and the second was the biochemical model already developed to use *K*D values to predict repression (Figure 6A). The training process simultaneously minimized the *K*D residuals for the 6 miRNAs with RBNS data and the repression residuals for 16 miRNAs with repression data (excluding let-7a to avoid the complications of the endogenously expressed let-7). The model was able to tune both the neural network weights and [*A*] values to minimize this joint objective function. Although the transfected duplexes were each designed to favor loading of the guide strand over the passenger strand, some repression of transcripts with canonical sites to the passenger strands was often detected (fig. S6A). Accordingly, the model considered both strands of each transfection experiment, which allowed the neural network to learn from another 17 potential miRNA sequences (with passenger strands with no detectable repression acting as negative training examples).

During training, we systematically left out all of the data for one of the eleven miRNA duplexes without RBNS data and trained on the data for the remaining ten, plus the data for the six duplexes with RBNS data, with the goal of predicting the 12-nt *K*D values for the two miRNAs of the held-out duplex, in an 11-fold cross-validation procedure. Testing these *K*D values on each held-out dataset reveal strong predictive performance (Fig. 6B, fig S6B), which exceeded that of TargetScan (Fig. 6C). To avoid using any data from the test set for any type of training, the average guide and passenger-strand [*A*] values fit for the training miRNAs were used to predict repression for the test miRNA (Fig. 6B). As expected, performance for most miRNAs improved if we use the average guide and passenger strand concentrations for that miRNA from the other cross validations. (Supplementary Figure 5). Sentence describing the only-canonical-site analysis (Fig. 6C) and whether the model is somehow finding non-canonical sites.

Overall, when using the *K*D values predicted from the neural network, the performance of the biochemical model exceeded that of Targetscan (Fig. 6D). More surprisingly, this performance rivaled that of the biochemical model applied to experimentally determined *K*D values (Fig. 5). Presumably the ability to account for the effects of both strands of the miRNA duplex largely offset the reduced ability to account for noncanonical targeting and any other disadvantages of using less accurate *K*D values.

Interestingly, the model is not able to generalize as well to some miRNAs, particularly miR-182 and miR-223, as to the others. This is presumably because these miRNAs occupy a more sparsely populated area of the possible sequence space, though there do not appear to be any obvious sequence differences between these miRNAs and the rest of the miRNAs. The model is also not able to out-perform TargetScan7 (Figure 6D) in predicting de-repression of miR-122 targets in miR-122 knockout liver cells from mouse (Eichhorn et al., 2014). In addition, because the model is trained on only endogenous miRNAs, which have a non-uniform distribution of dinucleotide frequencies, we are unsure how well the model would generalize to predicting repression of synthetic shRNAs. For example, none of the miRNA guide strands we considered have a CG dinucleotide in their seed sequences. For these reasons, we believe that our modeling approach is promising, but requires more training data to reliably be able to predict binding and repression for arbitrary miRNAs.

**Insights into miRNA targeting**

Our results provide new insight into both the canonical and non-canonical miRNA site types. For each miRNA, the canonical 8mer was the highest-affinity site identified, illustrating its primacy in miRNA targeting, and for miR-1 and let-7a, the next three most effective site types were the canonical 7mer-m8, 7mer-A1 and 6mer sites. However, the other four miRNAs each had 1–2 non-canonical sites with affinities at least that of the 7mer-m8, as well as another 4–15 non-canonical sites with affinities exceeding that of the canonical 6mer site. The observation that canonical sites are not necessarily those with the highest affinity raises the questions of how canonical sites are distinguished from noncanonical ones and whether making such a distinction is of any use. Our results show that two criteria readily distinguished canonical sites from noncanonical ones. First, the six canonical site types were the only ones identified for most miRNAs, whereas the noncanonical site types were typically identified for only one miRNA, and never for more than two. Second, the four highest-affinity canonical sites occupy most of the specifically bound AGO2, even for miR-7 and miR-124, which have the largest and highest-affinity repertoire of noncanonical sites (Fig. 1F, Fig. 2). This greater role for canonical sites is presumably because perfect pairing to the seed region is the most efficient way to bind the silencing complex. Thus, to achieve equivalent affinity, the non-canonical sites must be longer and are therefore less abundant. For example, although the miR-124 7mer-m8 site has lower affinity than a 10-nt noncanonical site, the canonical 7-nt site occupies much more AGO2–miR-124 because of its 64-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(Lewis2005) and justifying the continued distinction between canonical and noncanonical site types.

The success of our biochemical model in predicting how each mRNA would respond to the transfected miRNA supports the conclusion that site-binding affinity is the major determinate of miRNA-mediated repression and that noncanonical sites measurably contribute to this repression in the cell. When developing our model, we also considered an alternative version in which the probability of binding at least one silencing complex determined the extent of repression (Fig. Sx). However, the current model was more predictive for repression, which supports the idea that each additional bound AGO molecule contributes additional repression. In the framework of our model, the global fit of 0.99 to parameter *b* suggests that the binding of one additional silencing complex to the typical mRNA approximately doubles its decay rate, thereby halving its abundance. In the concentration regime of our transfection experiments, this binding of additional Ago can be achieved with either a single 8mer site with favorable AU context or two average 7mer-m8 sites.

The other parameter fit to each miRNA was [*A*], i.e., the concentration of loaded, unbound AGO in the cell.We note that because [*A*] is affected by the abundance of target sites in the cell, it was not expected to perfectly reflect the total concentration of loaded AGO. Indeed, we observed the expected relationship between the fitted [*A*] and predicted target-site abundance within the transcriptome(Denzler 2016) (fig. Sx). This relationship suggested that for these miRNAs the variance in our transfection efficiencies was small compared to the variance in available target sites and that our model is able to indirectly incorporate information about target-site abundance, a known correlate of miRNA-targeting efficacy(Garcia et al., 2011)(REF).

Some other features known to correlate with targeting efficacy are also incorporated in our biochemical model. Indeed, the contribution of certain features, such as site type(Bartel2009), predicted seed-pairing stability(Garcia et al., 2011), and nucleotide identities at specific miRNA/site positions(Agarwal2015), are expected to be more accurately captured in the miRNA-specific *K*D values of the 12-nt sequences than when generalized across miRNAs, which presumably helps explain the superior performance of the biochemical model. At the other extreme, the contributions of other features′-(Grimson2007), sitedistance between the site and the end of 3′ UTR(Grimson2007), are not at all captured the biochemical model. Thus, incorporating these additional features is expected to, especially when supplemented with RBNS-determined affinities of additional miRNAs

**FIGURE LEGENDS**

**Figure 1: AGO–RBNS enables quantitative analysis of relative binding affinities of canonical and novel miR-1 target site­–types.**

**(A)** Canonical sites of miR-1. These sites include varying amounts of contiguous pairing through the miRNA seed region (nucleotide positions 2–8), and may additionally contain an A nucleotide opposite miRNA nucleotide position 1. The single 8- and two 7-nt sites (the 8mer, 7mer-m8, and 7mer-A1), which have complementarity to the entire miRNA seed (red), are the most effective and most well-conserved sites, while the three 6-nt sites (the 6mer, 6mer-m8 or offset 6mer, and 6mer-A1) are less effective and less preferentially conserved throughout mammalian 3′ UTRs.

**(B)** AGO-RBNS. Purified AGO2–miR-1 is incubated with excess RNA library containing 37 randomized nucleotides until binding equilibrium (left). The AGO2–miR-1–bound library molecules are isolated by applying the reaction to a nitrocellulose filter under vacuum (middle), and then used to generate DNA amplicons for RNA-Seq (right). Five binding reaction spanning a 100-fold range in AGO2–miR-1 concentration and a mock binding reaction with no AGO2–miR-1 were sequenced, as well as a sample generated directly from the input RNA library.

**(C)** Fractional abundance of reads containing either of the six canonical miR-1 site-types, or lacking a site, in the 7.6 pM AGO2–miR-1 library plotted against their respective abundance in the input library. Dashed vertical lines depict the enrichment or depletion of each category in the AGO2-miR-1-bound library over the input library, and the dashed diagonal line represents x = y. Reads containing multiple distinct site types were assigned to the site type category with the greatest overall enrichment.

**(D)** Enrichment of reads containing either of the six canonical miR-1 site-types, or lacking a site, for each of the five AGO2–miR-1 concentrations across within the AGO-RBNS experiment, in comparison to the input RNA library. Colors correspond to the site type–categories described in (C). Points represent the experimentally determined enrichment values from the RBNS experiment, while the lines represent simulated enrichment value produced from a mathematical model of the experiment. (Inset) Relative dissociation constant (*K*D) values obtained from the mathematical model by maximum likelihood estimation (MLE). Reported for each measurement is the geometric mean value ± the 95% confidence interval given by resampling the read data, removing one AGO-miR-1 concentration sample from the data, and fitting the model to the remaining four samples 200 times.

**(E)** Experimentally generated (points) and model-simulated (lines) enrichment of an expanded, *de novo* set of site types identified by kmer analysis of enriched motifs observed in AGO-RBNS experiment with AGO2-miR-1. Site type–categories present in (B–D**)** have consistent colors. The ordering of legend represents the ranking of each site type by its enrichment value.

**(F)** *K*D values estimated from the mathematical model for each of the sites in the *de novo* set. Each site is classified by whether it contains a 7–8-nt canonical site (purple), a 6-nt canonical site (cyan), a noncanonical site (pink), or a sequence motif with no clear complementarity to miR-1 (gray). The point and error bars corresponding to each site type represent the geometric mean value ± the 95% confidence interval when resampling the read data, removing one AGO-miR-1 concentration sample from the data, and fitting the model to the remaining four samples 200 times.

**(G)** (Left) The fractional abundance of each site-type among all RNA library molecules bound by AGO2–miR-1 as inferred by the mathematical model, plotted over a range of AGO2–miR-1 concentrations encompassing the five sequenced experimental samples. The area within each segment corresponds to the fraction of that site type within the model at that AGO2-miR-1 concentration. Colors are as in (B). (Right) Pairing diagram as in (A), describing the pattern of pairing (blue), wobble pairing (cyan), mismatched pairing (red), bulged nucleotides (compressed rendering of two adjacent nucleotides), and end non-complementarity (gray; B = C, G, or U; D = A, G, or U; H = A, C, or U; and V = A, C, or G) in the definition of each of the additional sites identified by *de novo* kmer analysis of the AGO2–miR-1 high-throughput binding data.

**Figure 2: Extension of AGO-RBNS to let-7a, miR-155, miR-124, lsy-6, and miR-7 enables quantitative measurement of miRNA–specific binding preferences.**

**(A–I)** Relative *K*D values estimated by performing AGO-RBNS and subsequent mathematical modeling on purified AGO2–let-7a (A), –miR-155 (C), –miR-124 (E), –lsy-6 (G), and –miR-7 (I). Each site type is classified according to whether it contains a 7–8-nt canonical site (purple), a 6-nt canonical site (cyan), an enhanced 6-nt canonical site harboring additional complementarity to positions 7 and 8 separated by a bugled nucleotide (blue), a noncanonical site with partial complementarity to the seed region (pink), a site with complementarity to the 3′ region of the miRNA sequence (green), or a sequence motif lacking clear complementarity to miR-1 sequence (gray). The point and error bars corresponding to each site type represent the geometric mean value ± the 95% confidence interval when resampling the read data, removing one AGO-miR-1 concentration sample from the data, and fitting the model to the remaining four samples 200 times.

**Figure 3: AGO-RBNS–derived binding affinity data provide quantitative insight into the remodeling of miRNA–target RNA binding energetics by Argonaute, and correlate with in vivo repression efficacy.**

**(A)** Positional relative *K*D values for each potential 11mer motif spanning all six AGO-RBNS experiments in this study. The 8mer, 6mer, and 6mer-m8 *K*D values are included as reference. The solid lines denote the mean *K*D value across each miRNA sequence, and the left- and right-hand dashed lines denote the 95% confidence intervals obtained by resampling the read data, removing one AGO-miRNA concentration sample from the data, and fitting the model to the remaining four samples 200 times.

**(B)** (Left) Thermodynamic cycle depicting the contribution of the A1 and m8 sequence features as independent features in establishing the binding differences between the 6mer, 7mer-A1, 7mer-m8 and 8mer site types, for all miRNAs. (Right) The apparent binding contribution (∆∆*G*) of the A1 (blue and cyan bars) or m8 (red and pink bars) sequence features, determined from the ratio of relative *K*D valueof the 7mer-A1 to that of the 6mer (blue), the8mer to that of the 7mer-m8 (cyan), the 7mer-m8 to that of the 6mer (red), or the 8mer to that of the 7mer-A1 (pink), for all six AGO2-miRNA complexes analyzed. The coefficient of variation reports on the degree of similarity of ∆∆*G* of the A1 and m8 sequence features given by the 8mer/7mer-m8 and 8mer/7mer-A1 relative *K*D ratios, respectively, with that given by the 7mer-m8/6mer and 7mer-A1/6mer relative *K*D ratios, respectively, across all six miRNAs.

**(C)** Relative *K*D values of the 6mer (filled circles) and 7mer-m8 (hollow squares) plotted against their corresponding ∆*G* values as predicted by nearest-neighbor rules of RNA–RNA binding energetics, for all six AGO-miRNA complexes analyzed by AGO-RBNS. The solid and dashed lines depict the fitted relationship between observed *K*D and ∆*G* for the 6mer and 7mer-m8 site types, respectively, across all six miRNAs. The gray line depicts the expected relationship given by the fundamental thermodynamic equation ∆*G* = −*RT* ln *K*.

**(D–I)** The relationship between site-type AGO-RBNS derived relative *K*D and transcriptome-wide site-type repression efficacy for miR-1 (D), let-7a (E), miR-155 (F), miR-124 (G), lsy-6 (H), and miR-7 (I). The y-axis values and error bars denote the regression coefficient and associated standard error for each site type when modeling the log2(fold-change) of each mRNA as a linear contribution of each site-type present within 3′ UTR of that mRNA. The x-axis values and error bars are as in Figure 1(F) and Figure 2. Each dashed line represents the apparent relationship between binding affinity and efficacy of repression, determined by weighted-least squares regression using the pairwise ln(relative *K*D) values and regression coefficients, along with their respective standard errors. The coefficients of variation are weighted Pearson correlation derived from the same pairwise values.

**Figure 4: Flanking dinucleotide sequence context causes 100–fold differences in binding affinity across miRNA sequence and target site types, due to variation in target-site accessibility.**

**(A)** Enrichment of AGO2–miR-1 site-types across the five AGO-RBNS samples (points), and the corresponding model-simluated (lines), with the 8mer–containing reads in each sample further separated into 256 categories according to the identity of the 5′ and 3′ dinucleotides immediately flanking the site. Each non-gray point and line combination represents a single 8mer–flanking dinucleotide category, colorized such that AU-rich flanking sequence categories are preferentially blue, and GC-rich flanking sequences are preferentially green. The gray point and line combinations represent the non-8mer site–type categories depicted in (1D).

**(B)** Relative *K*D values for each site type–flanking dinucleotide combination flanking dinucleotide, across all miR-1 site types identified in (1E). The relative *K*D values within each row are obtained by separating only the reads corresponding to that site type, and reapplying the mathematical model as depicted in (A). Each point represents the geometric mean value obtained by resampling the data, removing one AGO-miR-1 concentration sample from the data, and fitting the model to the remaining four samples, 200 times. The points are colorized as in (A).

**(C)** Consistency of flanking dinucleotide effect across miRNA and site type. (Left) Comparison of predicted relative *K*D values to those obtained in (B), when aggregating all flanking dinucleotide contexts and canonical site types (the 8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-m8 and 6mer-A1), across all six miRNA studied. The predicted relative *K*D valueswhere given by linear regression, when modeling ln(relative *K*D) as the sum of a miRNA–site type term (with 36 distinct possibilities), and four additional terms corresponding to the identity of each of the four flanking dinucleotide positions (with 4 possibilities for each term), two more terms describing allowing the nonlinear interaction between the two 5′ nucleotides and between the two 3′ nucleotides. The coefficient of variation quantifies the degree of agreement between the predicted and actual ln(relative *K*D­) values. The flanking dinucleotide–related terms in the model used to predict the relative *K*D values corresponding to each miRNA were determined by linear regression on the other five miRNAs. (Right) The coefficients corresponding to each of the four nucleotides at each of the four nucleotide positions,

**(D)** Relationship between the relative *K*D of each miR-1 8mer–flanking dinucleotide combination and the averaged structural accessibility of the target sequence window corresponding to miRNA nucleotides 1–15 over all reads containing that 8mer-flanking dinucleotide combination within the sequenced input library, across all 256 flanking dinucleotide 8mer contexts. The coefficient of variation and the dashed line are determined by Pearson correlation and linear regression, respectively, with the pairwise ln(relative *K*D) and ln(average accessibility).

**(E)** (Left) The cumulative distribution of structural accessibility of the target sequence window corresponding to miRNA nucleotides 1–15 over all miR-1 8mer–containing reads in the input library (solid black line), the 7.2 nM AGO2–miR-1 sample library (solid red line), and when sampling reads from the input library to match the flanking dinucleotide frequencies of the 8mer-containing reads in the 720 nM AGO2–miR-1 sample (dashed red line). The central point within each line represents the geometric mean of the corresponding distribution, with the effect given by the difference in the geometric mean of the sampled distribution to that of the input, normalized to the difference in the geometric means of the 7.2 nM AGO2-miR-1 sample library to that of the input. (Right) The increase in structural accessibility of miR-1 8mer–containing reads in the input library when resampled to match the flanking dinucleotide content in each of the five AGO2-miR-1 sample libraries. For each sample library, the fractional effect is given by the difference in the geometric mean of distribution from the resampled input reads to that of the full distribution of input, normalized to the difference in the geometric mean of the distributions of the AGO-miR-1 sample library to that of the input. The left-hand most bar depicts the effect highlighted with the cumulative distribution plots (left).

# Supplemental Figure legends

**Figure S1. Reproducibility of AGO-RBNS–derived Data and Analysis; related to Figures 1 & 2.**

**(A)** Pairwise comparison of enrichment of *k*mer of length 9, between two independent AGO–RBNS binding reactions and library preparations. Additionally, the two binding experiment each contain a distinct preparation of purified AGO2–miR-1, as well as two different RNA pool, each generated from a distinct synthetic DNA containing and containing a distinct 5′ constant sequence. Plotted are all 8mer (purple), 7mer-m8 (red), 7mer-A1 (blue), 6mer (cyan), 6mer-m8 (violet), and 6mer-A1 (light blue) –containing *k*mers, as well as 10,000 arbitrarily chosen *k*mers not containing any of these sites (gray). The coefficient of variation is given by the Pearson correlation coefficient of the log-transformed data pairs. The dashed line indicates as reference x = y.

**(B)** Pairwise comparison of the biochemical model parameter estimates across the least-well-correlated-sample pair, when repeating the MLE procedure five times, excluding each of the five AGO2–miR-1 bound samples from the likelihood fitting. The parameters specify a *K*D for each of the site types, the stock concentration of the AGO2–miR-1 complex (light gray), and the equivalent concentration of AGO2–miR-1–independent library RNA nonspecifically copurifying with the AGO2–miR-1-bound library RNA (dark gray). The coefficient of variation is given by the Pearson correlation coefficient of the model parameters. The dashed line indicates as reference x = y.

**(C)** Table of all pairwise coefficients of variation when repeating the MLE procedure five times, excluding each of the five AGO2–miR-1 bound samples from the likelihood fitting. The bolded coefficient represents the pairwise comparison shown in(B).

**(D–I)** Position-specific enrichment of canonical sites for miR-1 (D), let-7a (E), miR-155 (F), miR-124 (G), lsy-6 (H), and miR-7 (I), across 25 possible registers within the randomized RNA library, counting from the 5′ end. Enrichment values are calculated using the most-concentrated AGO2–miRNA–bound sample corresponding to the AGO–RBNS experiment performed for that miRNA.

**(J–L)** Position-specific enrichment of 3′ autonomous site types for miR-155 (J), miR-124 (H), and lsy-6 (L), across 25 possible registers within the randomized RNA library, counting from the 5′ end. Enrichment values are calculated using the most-concentrated AGO2–miRNA–bound sample corresponding to that miRNA.

**Figure S2. Analysis of canonical, noncanonical, and context-dependent noncanonical site types (CNST) identified in Kim, Sung, Park *et al*., 2016; related to Figure 2.**

**(A–F)** Relative *K*D values estimated by performing AGO-RBNS and subsequent mathematical modeling on purified AGO2–miR-1(A), let-7a (B), –miR-155 (C), –miR-124 (D), –lsy-6 (E), and –miR-7 (F), substituting the site type definitions contained within this study for those proposed in Kim, Sung, Park, *et al*., 2016. The 7mer-m3.9 corresponds to the offset 7mer from that study; CDNST 1 corresponds to a A nucleotide across from miRNA nucleotide position 1 and pairing to positions 2–6; CDNST 2 corresponds to an A nucleotide across from position 1, pairing to positions 2–4, a mismatch across from position 5, and resumed pairing to positions 6 and 7; CDNST 3 corresponds to a mismatch across from position 2, pairing to position 3, a mismatch across from position 4, pairing to position 5, a mismatch across from position 6, and resumed pairing to positions 7–9; and CDNST 4 corresponds to an A nucleotide across from position 1, pairing to positions 2–4, mismatches at positions 5–7, and resumed pairing to position 8. For those CDNSTs with internal mismatches in their definition, all mismatched variants were combined for the model fitting, in order to produce one relative *K*D representative of the overall binding affinity of the site type. The point and error bars corresponding to each site type represent the geometric mean value ± the 95% confidence interval when resampling the read data, removing one AGO-miR-1 concentration sample from the data, and fitting the model to the remaining four samples 200 times.

**Figure S3. Analysis of the effects of bulged, pivot-bulged, and deleted target nucleotides on target site binding efficacy; related to Figure 2.**

**(A–F)** Relative *K*D values estimated by performing AGO-RBNS and subsequent mathematical modeling on purified AGO2–miR-1(A), let-7a (B), –miR-155 (C), –miR-124 (D), –lsy-6 (E), and –miR-7 (F), substituting the site type definitions contained within this study for a list including the six canonical site types (gray points), and as well as a series of site types spanning either all possible single-nucleotide bulges (left), or all possible single-nucleotide deletions (right). The point and error bars corresponding to each site type represent the geometric mean value ± the 95% confidence interval when resampling the read data, removing one AGO-miR-1 concentration sample from the data, and fitting the model to the remaining four samples 200 times. The dashed horizontal lines are provided as a visual aid. Bulged site types are classified as ambiguous (colored circles connected by transparent colored lines), in those cases where a nucleotide is added within a homopolymeric stretch within the site (i.e., the sequence ACAUUUCCA could be classified as a 8mer-bU4, 8mer-bU5, or 8mer-bU6 for miR-1), and the pivot-bulged site is defined as inChi *et al*., 2012, wherein there target nucleotide which pairs to miRNA nucleotide 6 is present twice (i.e., ACAAUUCCA for miR-1). Ambiguous deletion positions (black circles connected by vertical, transparently gray lines) arise from stretches of repeated nucleotides within a target site, such that removal of any of the nucleotides within the stretch produces the same deletion site.

**Figure S4. Extension of the analysis of flanking dinucleotide binding affinity to five additional miRNAs, and analysis of the relationship between structural accessibility and flanking dinucleotide binding affinity for the miR-1 8mer site type. Related to Figure 4.**

**(A–E)** Relative *K*D values for each site type–flanking dinucleotide combination flanking dinucleotide, across all site types identified for let-7a (A), miR-155 (B), miR-124 (C), lsy-6 (D), and miR-7 (E). The relative *K*D values within each row are obtained by separating only the reads corresponding to that site type, and reapplying the mathematical model as depicted in (A). Each point represents the geometric mean value obtained by resampling the data, removing one AGO2-miRNA concentration sample from the data, and fitting the model to the remaining four samples, 200 times. The points are colorized as in Figure 4.

**(F)** The cumulative distribution of structural accessibility of the target sequence window corresponding to miRNA nucleotides 1–15 over all miR-1 8mer–containing reads in the input (black), the 7.6 pM AGO2–miR-1 sample (pink), the 760 pM AGO2–miR-1 sample (purple) and the 760 pM AGO2–miR-1 sample (blue) libraries.

**(G)** The correlation between relative *K*D value for all 256 miR-1 8mer flanking dinucleotide combinations and the the geometric mean of target site structural accessibility distribution for the corresponding reads in the input library, varying both the length and the position of the target sequence window used for calculating site accessibility. Previous studies of siRNA effectiveness and miRNA targeting used a 15 nucleotide window centered between miRNA positions 7 and 8 (gray box, *r*2 = 0.80), while the greatest agreement in this study occurred with a10-nt window opposite miRNA positions 1–10 (black box, *r2* = 0.83).

**(H)** (Left) Pairwise comparison of the frequencies of all 256 8mer–flanking dinucleotide combinations for miR-1 in the input (x-axis) and 7.6 pM AGO2–miR-1 bound sample (y-axis), when the input is sampled to match structural accessibility distribution of the target sequence window corresponding to miRNA nucleotides 1–15 in the 7.6 pM AGO2–miR-1 bound sample. The coefficient of variation is determined by Pearson correlation of the log-transformed means of the sampled and observed values. (Right) The cumulative distribution of structural accessibility of the target sequence window corresponding to miRNA nucleotides 1–15 over all miR-1 8mer–containing reads in the input (solid black line), the 7.6 pM AGO2–miR-1 sample (solid red line), and in the sample of input reads used for the correlation analysis to the left.

**EXPERIMENTAL PROCEDURES**

**AGO–miRNA Complex Purification**

Human Embryonic Kidney 293 (HEK293) cells were transfected with pcDNA3.3 (Invitrogen) overexpression plasmid containing human AGO2 with an N-terminal 3XFLAG separated with a di-alanine spacer, using Lipofectamine 2000 (Thermo Fisher) in Opti-MEM (Gibco). 48 hours later, cytoplasmic S100 extract was prepared as previously described (Wee *et al.*, 2014), except that the lysate was cleared by passing it through a ?? gauge needle ~10 times. The S100 extract was flash frozen in ~1 mL aliquots and stored in liquid nitrogen. Specific AGO–miRNA complexes were prepared using a variation of a previously described method (Jasso–Flores *et al.,* 2013). Briefly, The S100 extract was loaded with 100 nM synthetic miRNA duplex for two hours at 20º C, and then the loading reaction was incubated for 30 minutes with magnetic Dynabeads (Invitrogen) that had been pre-immobilized with a biotinylated, 2′-O-methyl-RNA oligo containing a canonical 8mer site against the loaded miRNA. Each complex was eluted by incubating the magnetic beads with a synthetic DNA oligo complementary to the biotinylated oligo for one hour followed by magnetic separation. All buffers were as previously described, but with 1 mM Mg++ and no DTT. This eluate was then subjected to further purification with magnetic ANTI-FLAG beads (Sigma), as per the manufacturer protocol, but with a modified buffer containing 18 mM HEPES pH 7.4, 100 mM K+ acetate, 1 mM Mg++ acetate, 0.01 mg/ml yeast tRNA (Sigma). Upon elution of the AGO–miRNA complex via magnetic separation, DTT and glycerol were simultaneously added to a final concentration of 5 mM and 20% v./v., respectively. The synthetic miRNA duplex used to load the cytoplasmic lysate was prepared with a known proportion of 5′-radiolabeled guide, so that the final concentration of each AGO–miRNA sample could be quantified by autoradiography against a dilution series of the original duplex.

**Equilibrium AGO–RNA Bind-N-Seq**

Each equilibrium AGO-RBNS experiment comprised five binding reactions including 100 nM synthetic random library (IDT, see oligo list) and a variable concentration of a given AGO–miRNA complex, was performed by first preparing a concentration series of a given AGO­–miRNA complex, where the greatest concentration of the complex was 40% (v./v.) in the final reaction, and each of the four additional binding reactions in the series were diluted 3.16–fold, resulting in a 100–fold range of the complex over five samples. Additionally, we performed one mock binding reaction in which the randomized library was incubated with the protein storage buffer absent the AGO–miRNA complex. Since the stock concentration of the five AGO–miRNA complexes ranged from ????–???? nM, we set the concentration of the randomized library in all reactions to 100 nM, so that the AGO–miRNA complex–library molar ratio would vary between ???? and ????. Each individual binding reaction was 20 µL, with 18 mM HEPES pH 7.4, 100 mM K+ acetate, 1 mM Mg++ acetate, 0.01 mg/ml yeast tRNA (Sigma)

***De novo* site type identification from AGO-RBNS read count analysis**

A list of apparent site types was constructed for each of the six miRNAs for which AGO-RBNS was performed. This was done by calculating the enrichment of all 10mers in the sequencing data from the five AGO-miRNA bound libraries in comparison the input library. For a given miRNA, each 10mer was queried for complete complementarity to the miRNA sequence, 2.) near–complete complementarity to the miRNA to the miRNA seed sequence (nt 2–7 with an A across from miRNA position 1), or complementarity to another region of the miRNA spanning the entire length of the 10mer. In the absence of this, the 10mer was queried, in order for 1.) complementarity to the miRNA at 9 positions within the 10mer with an internal bulged nucleotide, 2.) complete complementarity to the miRNA at all 10 positions while allowing for wobble pairing 3.) complementarity to the miRNA at 9 positions of the 10mer with an internal non-bulged, non-wobble mismatch position, 4.) complementarity to the miRNA at 9 positions of the 10mer and a single internal bulged target nucleotide, while allowing wobble pairing, and 5.) complementarity to the miRNA at 8 positions within the 10mer, with both an internal bulged nucleotide and a mismatch position. Any identified pairing configurations are then stored, and the process is repeated on the two 9nt sub-kmers within the 10mer motif, the three 8mers, etc., until a kmer length is identified for which a single kmer is completely complementary to a region of the miRNA. The resulting list of candidate miRNA-target site types are then ranked via a scoring system with a set of weights that rewards each Watson–Crick paired position within the site (preferentially to nucleotides 2–8, 12–16, 17–22 or 23, and 9–11, in that order), each dinucleotide of Watson–Crick (uniformly across the miRNA sequence), the presence of a contiguous pairing to miRNA nucleotides 2–5, and A/U content outside of the internal region of the 10mer defined as participating in the miRNA–target interaction. The scoring system penalizes bulged nucloetides, wobble pairing, and mismatched pairs in that order, and G content outside of the internal region of the10mer defined as participating in the miRNA–target interaction. The weights associated with each feature were arbitrarily tuned such that the site type identified within each 10mer was consistent with visual inspection of the 10mer, and in addition that the majority of the top 10mers were identified as containing the same site type. Upon identification of a site type from analysis of the top 20 10mers, reads containing this site were removed from the sequencing data from both the input library and the five AGO–miRNA bound samples.

We note that while this approach is *ad hoc*, we find it suitable since it effects only the naming of these novel site types, and cannot influence their relative enrichments within the RBNS data, nor their estimated *K*D values. Secondarily, due to the variable amount of sequence redundancy within the miRNA guide sequences studied (i.e., miR-1: UGGAAUGUAAAGAAGUAUGUAU, let-7a: UGAGGUAGUAGGUUGUAUAGGU, etc.), knowledge of the true binding preference within the 10mer cannot be unambiguously known.