**TO DO LIST:**

1. Lsy-6 doesn’t have an enhanced 6mer site DONE
2. Systematically check the pivot-bulged sites. (Put them in supplemental Figure 2; right next to the Daeyun Beck figures. DONE
3. Check which bulge sites are in the top list DONE
4. Average the Daeyun Beck sites; don’t have them all iterated out. DONE
5. Figure legends.
6. Gradient math.
7. 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.
8. Figure out if double sites do better than single sites acting linearly.
9. Methods (math and experimental)
10. Ask people which they’d rather see: Model vs Data or Data vs model: keep it consistent.
11. Make a schematic for Figure 2 with the adding nucleotides, A1 and m8.
12. Add supplemental figure with the flanking Kds for the other four miRNAs.Delete the heatmap with the 6 canonical sites. DONE
13. Bootstrap for the linear model to show that the interaction terms are significant.
14. Make the pivot bulge sites brown-gold.
15. Plot off-rates against ∆∆G. (Namita)

**AFTER SUBMISSION:**

1. Repeat lsy-6 unbiased analysis
2. miR-155: Manual Kd for the 11mer?
3. Include figure showing that flanking kds don’t change as a function of window length?

**Title**

Sean E. McGeary, …, David P. Bartel

**ABSTRACT**

**HIGHLIGHTS**

**INTRODUCTION**

MicroRNAs (miRNA) comprise a class of 21–23 nt non-coding RNA that regulate gene expression at the post-transcriptional level{Bartel:2009fh}. Individual miRNA molecules achieve their functional form in association with an Argonaute (Ago) protein, wherein the miRNA directs the Ago–miRNA complex to specific transcripts, promoting mRNA destabilization and in some cases translational repression. The persistence of individual miRNA genes throughout diverse metazoan species, the large number of reported miRNA knockout phenotypes in fly and mouse (Dave’s table), and the finding the majority (60%) of mammalian mRNAs harbor at least one conserved miRNA site, together underscore the important role for these RNAs in the coherent regulation and maintenance of gene expression underlying biological life{Grimson:2008cv}{Friedman:2008km}. A more precise understanding of miRNA function, potentially providing insight into their evolutionary origin as well as their role in health and disease, requires an accurate model of the selectivity and magnitude of miRNA-mediated repression across the diverse set of mRNAs expressed in animal cells.

Evidence from computational, structural, biochemical, and organismal studies has repeatedly demonstrated the importance of pairing to the miRNA seed sequence (nt 2–7) for miRNA function{Lewis:2005cb}{Lai:2002dz}{Schirle:2012bf}{Wee:2012df}{Hafner:2010kr}. In particular, genome-wide experiments monitoring changes in mRNA, protein, and translation levels in response to either the induction or removal of a single miRNA have enabled the comparison of different extents of pairing for their effect on repression. These studies suggest that the most effective target sequences can be classified into one of four so-called canonical site-types of increasing effect: 1.) The minimal 6mer site, containing perfect complementarity to the miRNA seed, 2.) the 7mer-A1, with seed pairing and an A across from miRNA position 1, 3.) the 7mer-m8, with seed pairing and an additional pair to nt 8, and 4.) the 8mer with seed pairing and features of both 7mer site-types{Bartel:2009fh}. Two more site types with complementarity to the seed region have been subsequently described: the offset 6mer, with complementarity to nt 3–8, the 6mer-A1, with complementarity to nt 2–6 and an A across from position 1, due to detectable repression in global analyses of particular studies as described above, and evidence of their conservation(ref?).

While seed pairing predominates among examples of sites mediating effective repression, there is evidence that other portions of the miRNA guide sequence can contribute to miRNA function. Extensive pairing through the central region (nt 3 or 4—14 or 15) can mediate repression, without seed pairing (Shin?), and pairing to the 3′ end of the miRNA (nt 11–16), while not sufficient for independent regulation, can increase the effectiveness of either a site with canonical pairing or imperfect pairing to the 5′ end of the miRNA. A much greater number of non-canonical site types have been recently reported owing to their detection in data from high-throughput crosslinking-and-immunoprecipitation (CLIP) of Ago-miRNA complexes in vivo{Helwak:2013ga}{Loeb:2012bc}{Chi:2012jm}{Chi:2009ht}{Shin:2010hn}, and a variation of CLIP containing chimeric miRNA–target mRNA due to the presence of ligase activity during library preparation. These non-canonical sites vary in terms of both the region of the miRNA to which they are complementarity and also the extensiveness of their pairing, and include seed sites with internal wobble pairing, bulges, and mismatches unique to each miRNA sequence. It has been challenging to incorporate these non-canonical sites into a consistent model of miRNA action, due to both to the sparse, non-quantitative signal by which they were identified, as well as their lack of consistent effect in vivo.

While the four canonical site types together with the offset 6mer (herafter referred to as the 6mer-m8) and the 6mer-A1 are the only site types described which consistently exhibit repression, the best current efforts to predict in vivo expression changes remains surprisingly poor{Agarwal:2015bw}, with the coefficient of determination between prediction and actual data ranging between 0.001–0.14. Indeed, the lack of power across the wide range of target prediction programs, in addition to the range of heuristics employed to actually predict the effect of miRNAs on targets{Krek:2005er}{Gumienny:2015fs}{Breda:2015bv}{Agarwal:2015bw} demonstrates a fundamental limitation within miRNA research currently: the absence of a quantitative, mechanistic model of the pathway. This is on its face surprising, since the number of proteins that have been described as involved in miRNA–target interaction is modest: one of the few direct binding partner of a mature Ago–miRNA complex is the GW182 scaffold protein (TNRC6 in human), which itself binds to the host effector proteins mediating mRNA destabilization via deadenylation and decapping (REF?). Recent reports demonstrated that TNRC6 promotes phase-separated droplets that sequester AGO-miRNA complexes and target RNAs *in vitro,* which might suggest that miRNA mediated repression occurs in concentrated foci rather than as a diffuse phenomenon across the cytoplasm. In another recent study, a phosphatase *Akt1* and its corresponding kinase \_\_\_? Were identified as proteins contributing to effective miRNA mediated repression by the maintenance of a phosphorylation cycle acting on \_\_\_ residues of Ago. This would imply that the basal Ago-miRNA complex requires consistent molecular maintenance in order to persistently mediate its biological role. While both of these recent studies provide interesting insights into unexplored molecular complexities of miRNA mediated repression, and indeed of mechanisms of gene regulation in general, they provide no obvious solution to the generic problem of predicting which mRNAs will be most affected by a miRNA, since none of recent mechanistic complexity ascribed to either TNRC6 or *Akt1* is purported have sequence specificity.

To this end, the issues in constructing a likely, molecular mechanism of miRNA-mediated repression can be ascribed to two orthogonal unknowns: 1.) whether or not different Ago–miRNA complexes bind differently to their repertoire of cellular targets, and 2.) the extent to which all downstream steps in the molecular pathway of miRNA-mediated repression are agnostic to differences in miRNA–target binding. In order to begin to dissect the relationship between these two phenomena, there is a need for data of a kind not currently available with miRNA mediated repression: that of high-throughput, quantitative measurements which precisely measure the variation in binding affinity of individual Ago-miRNA complexes for all possible target sequences, for a number of miRNAs. We adapted RNA-Bind-N-Seq (RBNS) for use with purified AGO­–miRNA complexes{Lambert:2014jm}.

**RESULTS**

**Argonaute-RBNS (AGO–RBNS) yields quantitative binding affinity profile for canonical and previously unidentified target site types.**

RBNS enables the measurement of binding affinity of a particular RBP with an exhaustive list of putative binding sites, and is performed by incubating an RNA binding protein with a synthetic RNA library containing ~40 contiguous positions of random nucleotide composition until binding equilibrium, followed by high-throughput sequencing of both the protein-bound and input RNA library. The random library inherently contains a low frequency of canonical 6mer, 7mer, and 8mer sites within diverse sequence contexts, and a majority of RNAs sequences within with no complementarity, for any miRNA sequence. We reasoned that calculation of the overall binding affinity for each site-type within this randomized context would provide a more representative picture of binding for comparison with relative efficacy of repression in vivo than binding affinity.

We modified the published RBNS workflow by performing nitrocellulose filter binding to isolate the AGO–miRNA-bound library molecules, as opposed to using streptavidin-coated magnetic beads to isolate a strep-tagged variant of the protein{Lambert:2014jma} (Figure 1A). We favored filter binding for its rapid binding and wash steps in comparison to bead-based capture, reasoning this would improve retention of the real yet low-affinity binding events that may become depleted during the time required to wash the beads. Additionally, we reasoned that using nitrocellulose–nylon filter binding would make our measurements more directly comparable with the wealth of recent studies measuring the equilibrium binding and kinetics of specific AGO–miRNA complexes with individual target RNA sequences{Wee:2012df}{Schirle:2014cj}{Schirle:2015ct}.

We performed AGO-RBNS by incubating a constant concentration of the randomized library (100 nM) with five distinct concentrations of purified human AGO2–miR-1 (7.2–720 pM, logarithmically spaced). The experimental concentrations were chosen to set the ratio of total AGO2–miR-1 to total 8mer-containing library RNA in the binding reaction to be ~ 1:10 and 10:1 in the 7.2 and 720 pM samples, respectively, such that each of the five AGO-RBNS samples would report on a distinct state of saturation of the of the library sites. Indeed, comparison of the frequency of 8mer, 7mer, and 6mer–containing reads in the 7.2 pM AGO2–miR-1 sample to that of the input library sample demonstrates a clear enrichment of the four canonical site types (39.9, 15.5, 7.82, and 3.52 for the 8mer, 7mer-m8, 7mer-A1, and 6mer, respectively), and virtually no enrichment of either the 6mer-A1 or 6mer-m8 site types (1.02 and 0.97, respectively) (Figure 2B). This is consistent with reports of their typically weaker signal in meta analyses of repression{Friedman:2008km}{Agarwal:2015bw}{Kim:2016bo}.

Analysis of enrichment of the six seed sites across the five samples illustrates two hallmarks of this experimental platform: First, increasing the AGO2–miR-1 concentration from 7.2 pM to 22.8 pM and again to 72 pM leads to a monotonic increase in enrichment for site-types of all affinities, due to an increase in specific signal proportional the concentration of AGO2–miR-1 in relation to a constant, low amount of AGO–miRNA-independent library RNA recovered in the experiment. Second, in samples with greater than 72 pM AGO2–miR-1 exhibit we observe decreased enrichment of the 8mer site type, and in the 720 pM AGO2–miR-1 sample additionally observe decreased enrichment of the 7mer-m8 and 7mer-A1 site types, indicating the onset of saturation for these high-affinity site types within the random library{Lambert:2014jm}. By contrast, all three 6mers exhibit monotonic increases in enrichment across the entire concentration course (Figure 2C). These two features, AGO–miRNA-independent, non-specific binding and site type–saturation, illustrate that no individual experimental sample quantitatively reflects the relative binding affinity of these site types in through enrichment values. Indeed, currently published methodology, which incorporate neither background binding nor ligand saturation, produces a distinct set of relative *K*Dvalues for each RBNS sample, causing estimates of relative binding to be appear exacerbated or compressed depending on the which sample was analyzed. Application of this method therefore poses fundamental complications for mechanistic interpretation of differential binding affinity across miRNA and target sequences, and diminishes the prospect that these estimates could function as predictors of biological repression.

We created a novel computational strategy to calculate *K*D values that simultaneously incorporates information from all concentrations within an RBNS experiment. The model takes as input parameters specifying the (seven) *K*D values for each miRNA target site–type, and two additional parameters specifying the stock concentration of the purified AGO2–miR-1 complex and the constant amount background, AGO–miRNA-independent library RNA recovered in all samples, and outputs a prediction of the fraction of each site type within all five samples across the dilution series in the experiment. This model was used to perform maximum likelihood estimation (MLE) of all seven *K*D values in the experiment. The resultant model parameters explain the observed data strikingly well, as demonstrated by the similarity of enrichment values produced by the computational model and the experimental data (Figure 1C). Additionally, pairwise Pearson correlation of *K*D values obtained when excluding each concentration from the model fitting yielded no *r2* values below 0.999 (Data not shown), validating our analytical framework as well as the consistency of apparent binding affinities within the data.

We observe relative binding affinities across the six site types, with the 8mer, 7mer-m8, 7mer-A1, and 6mer site types exhibiting ~580, 210, 120, and 40–fold difference in *K*D in comparison to reads without a site. The 6mer-m8 and 6mer-A1 site types demonstrate ~4 and 6–fold higher affinity, in keeping with the inconsistent evidence of their efficacy across miRNAs and biological samples. Because our analytical approach to identify the site type *K*D values necessarily generates an underlying biochemical model, we looked at (change this) the proportion of each site in complex with AGO2–miR-1 within the model over the concentration course. We observe that the most commonly bound RNA molecules were those with no site type, ranging from ~35% to 60% over the concentration course. By contrast, the highest affinity 8mer site type ranged from ~18% to <5% of the bound RNA molecules. This result is qualitatively similar to reports showing that the effective number of target sites for a given miRNA contributed by the summed contribution of low affinity sites is of the same order of magnitude as canonical sites, and further supports the notion that a fundamental ceiling on the magnitude of miRNA-mediated repression is governed by the fact that an individual AGO-miRNA complex will spend as much as half its molecular life associated with a vast repertoire of low-affinity sites{Denzler:2014cz}{Denzler:2016eh}.

Considering the unprecedented precision of our approach for characterizing seed-region site binding preferences AGO2-miR-1, we sought to construct a *de novo* binding profile for AGO2–miR-1, for the purposes of discovering previously unknown miRNA target modes as well as their likely efficacy in comparison to known sites. To identify binding sites, we iteratively 1.) calculated the enrichment of all possible kmers for lengths 5–11 nt in the 72 pM AGO2–miR-1 bound library in comparison to the initial randomized pool, 2.) determined the kmer length for which the top-most enriched kmer exhibited the greatest relative enrichment compared to the second-most enriched kmer, and then 3.) designated this kmer as the motif, and removed all reads containing the motif from both the 72 pM and input libraries. We performed this iterative procedure until we were left with 11mer mofits with enrichment values < 2-fold. We then fit the binding model to this expanded list of site-types (Figure 1E, F). This approach demonstrated that the 8mer, 7mer-m8, and 7mer-A1, and 6mer site types were indeed still the highest affinity motifs, and also identified novel modes of miRNA–target engagement with binding affinity similar to that of the 6mer-m8 and the 6mer-A1. Comparison of these motifs to the sequence of miR-1 revealed that miR-1 can tolerate a wobble G or a mismatched U at position 6, a bulged U somewhere between positions 4 and 6, a mismatch at position 5, with affinity ranging from 5–10 fold above the remaining no-site reads. We also observed that the motif “ACACACA” demonstrated ~6–fold greater binding than background, which was surprising since its greatest extent of complementarity to miR-1 is at two UGU stretches at positions 6–8 and at 18–20. In aggregate, these results suggest that AGO2–miR-1 binds most effectively to the 8mer, 7mer-m8, 7mer-A1, and 6mer site types, but tolerates a diversity of binding modes of intermediate affinity. While alternative modes for miRNAs have been proposed based on studies using CLIP{Loeb:2012bc} and CLASH{Helwak:2013ga}, our approach allows for the direct quantification of the relative binding strength of these interactions, thereby allowing their coherent incorporation into a quantitative framework of predicted efficacy.

**Extension of AGO–RBNS analysis to let-7a, miR-155, miR-124, and lsy-6 reveals distinct binding modes across miRNAs.**

We extended our analysis to four more miRNAs, three of which (hsa-let-7a, hsa-miR-1, hsa-miR-155, hsa-miR-124,) were chosen for their conservation throughout vertebrate lineages and because of their recurrence throughout a variety of repression, CLIP, and *in vitro* studies, potentially enabling the comparison or synthesis of our findings with those of others. We also analyzed cel-lsy-6, the low-efficacy of which having previously been ascribed to poor seed pairing stability (SPS), a heuristic indicator of low binding affinity. We generated a purified AGO2-miRNA complex, performed AGO-RBNS, and measured the binding affinities with *de novo*, miRNA-specific binding motifs, yielding in total five distinct sets of miRNA–target RNA profiles of binding affinity (Figure 1F, 2A–D).

The binding affinity profile of let-7a (Figure 2A), demonstrates similar binding to miR-1, but with the 6mer-m8 and 6mer-A1 site types exhibiting greater binding affinity than the large majority of noncanonical sites identified. The noncanonical binding modes also correspond to wobble-pairing, bulged nucleotides, and mismatch positions with the seed, although to distinct nucleotide positions than that of miR-1, and, in the case of mismatches, different nucleotide identities.

Analysis of miR-155 (Figure 2B), miR-124 (Figure 2C), and lsy-6 (Figure 2D) revealed a distinct binding mode, exhibiting extended complementarity (9–11 nt) to the miRNA 3′ end. To corroborate the apparently dichotomous binding capacity of the miRNA 3′ end, we reanalyzed all five miRNAs, reassigning the reads based on the occurrence of any of the six seed sites, or of an 11nt window of complementarity to the miRNA sequence from guide position 3 to the 3′ end of the guide (Figure S2iA). We observed that indeed for miR-155, miR-124, and lys-6, pairing of 11 contiguous nt in the 3′ end exhibited binding affinity greater than that of the canonical 6mer site, while for let-7a and miR-1, all 3′ windows exhibited weaker binding affinity than the canonical 6mer. Within the *de novo* site list, the 3′ 11mer exhibited binding in excess of the 7mer-m8 site type corresponding to miR-155, miR-124, and lsy-6, suggesting that 3′ pairing modes might have significant regulatory potency within the transcriptome, but for a select subset of miRNAs. The 3′ binding of miR-155 is further distinguished from that of miR-124 and lsy-6, as it exhibits pairing across from nt 13–23, rather than nt 9–19. These results point to a role for miRNA guide length in remodeling the magnitude and guide position of 3′ end, as miR-155 is 1 nt longer than the other four miRNAs.

We decided next to analyze the five AGO–RBNS experiments with respect to a recent published report that posited an expanded set of target site–types, extending the putative scope of miRNA–mRNA regulatory interactions{Kim:2016bo}. This study confirms the 6mer-m8 and 6mer-A1 site types, and identifies the “offset 7mer” site type with pairing from nt 3–9, as well as four context-dependent noncanonical site types (CDNST). Upon assigning *K*D values according to these site type–categories (and not the *de novo* list previously described), we find no strong evidence for the offset 7mer site type, as the relative binding difference between it and the 6mer-m8 ranged between 0.52 and 1.45–fold (Figure S2iB–F). In addition, we find that the CDNSTs do not constitute classes of site types with broad predictive power: while CDNST 1 (equivalent to the 5mer-m2.6) with let-7a exhibits 8.7–fold greater affinity than that of no site–containing reads, all other CDNST-and-miRNA combinations yielded values between 0.85 and 3.77 above no site–containing reads.

We inspected the variety of single-nt wobble, bulge, and mismatched site types identified for miR-155, miR-124, and lys-6. Two of the bulge-containing miR-124 site types recapitulate prior observation of bulge-pivot pairing (the 8mer-bG(6.7) and 7mer-m8bG(6.7) site types){Chi:2012jm}, as well as two such sites for lsy-6 (the 8mer-bA(6.7) and the 8mer-bG(7)). To confirm this, we reanalyzed all five miRNAs with respect to bulged nucleotides (Figure S2iiA–E) and for sites that may have been missed in our *de novo* site types analysis. We confirm that only miR-124 and lsy-6 exhibit bulged-pivot site types with binding affinities in excess of their 6mer site, with let-7a the moderate binding affinity of the let-7a bulged-pivot site being mostly due to the appreciable binding affinity of its 6mer-A1 site type. In addition we find, as expected, that single-nucleotide deletions within the seed region (Figures S2iiF–J) are sufficient to eliminate the majority of the 8mer binding affinity. miR-124 exhibited two more distinct classes of site types, the first being representing canonical 6mer and 7mer-A1 sites with added binding efficacy due to either a bulged U at position 7 (8mer-bU(7.8) and 7mer-m8bU(7.8)) or a bulged A at position 8 (8mer-bA8), and the second consisting of a partial seed match with complementarity through to position 8 and two A nucleotides across from miRNA positions 9 and 10 (AA–8mer-bU6, AA–7mer-m8bU6, AA–6mer-m8 and AA–5mer-m8). Finally, miR-155 and lsy-6 exhibit five and three binding motifs with ambiguous miRNA–target recognition, while miR-124 only exhibited site types with discernable patterns with respect to the guide sequence.

We observe striking differences in the relative binding affinity between 8mer, 7mer, and 6mer sites across the miRNAs in this study: the difference in binding affinity between the 7mer-m8 and the 7mer-A1 site types for miR-1, let-7a, miR-155, and miR-124 and lsy-6 is 1.8, 3.5, 1.4, and 11.4, and 1.9–fold, respectively. Indeed, while prior studies have demonstrated that AGO proteins remodel the thermodynamic properties of their loaded RNA guides{Wee:2012df}{Salomon:2015fb}, the data presented here suggest that the nature of this remodeling is influenced by both the length and sequence content of the guide, and contributes to differences in relative affinities across canonical site types, in addition to influencing the distinct repertoire of noncanonical site types for that miRNA.

We probed the extent to which the overall efficacy in binding could be predicted by the intrinsic sequence features of each guide, by comparing the 6mer and 7mer-m8 *K*D values for each miRNA to its corresponding ∆GNN, as predicted by nearest-neighbor rules using RNAduplex (see Methods) (Figure 2E) (REF). We chose these two site types as they do not include the A1 sequence feature, as binding to this feature is not mediated by the miRNA itself. While a general trend is apparent for both site types (*r2* = 0.56 and 0.87), it is significant for only for the 7mer-m8 site type (*p =* 0.18 and 0.022, respectively), highlighting the limitation of predicted seed-pairing stability in evaluating the relative efficacy of different miRNAs for the same site type in global analyses of miRNA-mediated repression.

We next sought to study the relationship between the A1 and m8 sequence features. We find that these two features constitute a linear, thermodynamic cycle, as the product of the *K*D,rel for the 7mer-A1 and 7mer-m8 together is absolutely predictive of the *K*D,rel for the 8mer, across all five miRNAs (*r*2 = 1.00, *p* = 1.04 x 10–4) (Figure 2F). This suggests that in understanding the relationship between the four canonical site types for any miRNA, that three values are required: the binding affinity of the 6mer site, the relative improvement in binding due to the A1 sequence feature, and the relative improvement in binding due to the m8 sequence feature. We next looked at the relationship between the 6mer-normalized predicted ∆∆GNN values and the AGO-RBNS–derivied ∆∆G values. Surprisingly, we see that seed-region relative binding across all five miRNAs is well explained using the predicted ∆∆GNN values (*r*2 = 0.966, 0.969, 0.999, 0.993, and 0.8346, *p* = 0.0174, 0.0155, 0.0005, 0.0004, and 0.0864, for miR-1, let-7a, miR-155, miR-124, and lsy-6, respectively) albeit with a miRNA-specific scaling factor (Figure 2F). For let-7a and miR-124, this scaling factor is close to 1, suggesting the relationship behaves according to the thermodynamic expectation of *K*D,rel = e–∆∆G/RT; however, for miR-1, miR-155 and lsy-6 the relative *K*D for each of the three site types was exacerbated 1.8, 5.9, and 8.0–fold, respectively. These results indicate that miRNA loading seems to regularize the magnitude of the contributions of the A1 and m8 sequence features, such that guide sequences with shallow intrinsic binding differences in the seed region can still distinguish the four canonical site types in the context of the AGO–miRNA complex. We find this result especially surprising with respect to the energetic contribution of the A1 position; since this target nucleotide directly interacts with Ago protein, and thus is not expected to behave according to the energetics of the equivalent duplex in solution. Taken together, these results demonstrate that intrinsinc binding properties of individual AGO–miRNA vary with respect to each other over several properties, and that these differences are not easily predicted from standard thermodynamic models, engendering the need for highthroughput *in vitro* techniques for their direct measurement.

**Flanking dinucleotide sequence context causes 100–fold differences in site type binding affinity.**

It is well established that additional mRNA sequence features extrinsic to site type play a role in modulating the efficacy of repression, such as AU-content local to the target site, target site­–accessibility (due to low propensity to form secondary structures occluding the site), 3′ UTR length, and target site position within the 3′ UTR {Agarwal:2015bw}{Hausser:2009cn}. The identification of these and other sequence- and transcript-level features as contributing to miRNA-mediated repression occurs primarily through the detection of a correlation with that feature and the magnitude of repression, in studies measuring global changes in gene expression upon transfection or knockout of a particular miRNA sequence. Such studies are of limited utility for precise quantification of the relative effects of each of these features, however, because the repertoire of target sites against any particular miRNA within the endogenous transcriptome constitute a sparse cohort in comparison to the number of sequence features that have been identified. We decided to analyze these effects within the AGO-RBNS equilibrium experiments, reasoning that our complex biochemical data and computational modeling approaches would enable us to monitor the effects of sequence context at the level of miRNA-target binding with an unprecedented level of quantitative detail. Furthermore, we reasoned that *in silico* structural folding algorithms might provide more accurate structural accessibility predictions for the target sites within the randomized library in comparison to those occurring naturally within transcriptomic 3′ UTRs, which are too long to be computationally folded in their entirety.

In order to directly quantify the effects of sequence context on miRNA–target site binding affinity, we reclassified the 8mer-containing read counts within all the samples from the miR-1 equilibrium AGO-RBNS experiment, assigning each read to one of 256 possible categories according to the dinucleotide sequence immediately flanking either side of the site (e.g., AA–8mer–AA, AA–8mer–AC, …, UU-8mer-UU), and calculated *K*D values for each by repeating the modeling described earlier with an updated site list in which the single “8mer” category was replaced with the 256 8mer site-context (Figure 3A). This analysis reported a 100–fold range in 8mer *K*D (0.44 and 48.3 pM for AT–8mer–TA and GG–8mer–GG, respectively), with binding affinity across the flanking sequences tracking AU primary nucleotide sequence content. Extension of this analysis across all miR-1 site types (Figure 3B), and additionally to all site types for let-7a, miR-155, miR-124, and lsy-6 (Figure S3A–E) yielded a similar effect size and primary nucleotide trend. We find the magnitude of this difference striking, especially when considered in comparison to the binding affinity differences observed between miRNA-site types (e.g. 13-fold between the 8mer and 6mer site types for miR-1). We analyzed the generic features of the flanking dinucleotide effect across guide sequence and site type by fitting a linear model to the log-transformed *K*D values as a function of miRNA–site type combination, and primary nucleotide identity at each of the four positions within the flanking dinucleotide sequence, using data from the six seed sites of each miRNA, a cohort of 5,360 *K*D measurements (Figure S3E). The linear model (*r*2 = 0.94) indicates an independent contribution by the 5′ and 3′ dinucleotide sequences to binding affinity, with the 5′ flanking dinucleotide, which is spatially proximal to the central nucleotide binding channel of the Ago protein, contributing a ~2-fold greater effect on binding affinity than that of the 3′ flanking sequence, as per the magnitude of the nucleotide coefficients ascribed to each position by the model (Figure S3F). In all cases, A and U nucleotides both contribute positively to binding affinity, G nucleotides contribute negatively, and C nucleotides provide an intermediate or neutral contribution to binding affinity (Figure S3F). This result is in ready agreement with the finding that AU-rich sequence context contributes positively miRNA-mediated repression, and extends this rationale with the discovery that C and G nucleotides are not equally detrimental to target binding.

We sought to study the extent to which the range in flanking dinucleotide binding affinity could be explained by differences in the structural accessibility across the diverse set of targets within the random RNA library for each flanking dinucleotide category. We calculated the target site accessibility, defined as the probability that the region of the target RNA across from miRNA nt 1–15 is unpaired as calculated by RNAplfold, for each miR-1 8mer–containing read sequenced in the input RNA library. We chose the window across from miRNA positions 1–15 for consistency with studies of the effects of target accessibility on global siRNA efficacy (reference) and miRNA–mediated repression (Agarwal). We observe a distribution of values with geometric mean 2.04 x 10-3 for the 8mer-containing reads in the input library, and distributions with increased accessibility in the AGO2–miR-1 RBNS samples, with geometric mean values of 2.19 x 10-2, 2.30 x 10-2, 2.02 x 10-2, 1.52 x 10-2, and 1.07 x 10-2, for the 7.2, 22.8, 72, 228, and 720 pM samples, respectively, consistent with a model in which the most accessible target sites are preferentially bound in conditions of limiting Ago-miRNA complex, and that more suboptimal target sites become increasingly occupied with increasing Ago-miRNA complex in the binding reaction. That the distribution of values generated by RNAplfold exhibited these two features engendered our confidence that the algorithm reports on real differences across the target RNAs in our random library, for the purposes of interrogating the causal relationship between target site accessibility and the influence of flanking dinucleotide context on site type binding affinity.

We binned each 8mer-containing read based on the 256 possible flanking dinucleotide categories, and calculated the geometric mean target site accessibility for each category. We observed a striking correlation between target site accessibility and the *K*D value as measured within the RBNS experiment (*r*2 = 0.80, *p* < 2.2 x 10–16) (Figure 3C). Furthermore, all 16 flanking dinucleotide sequences containing a GG dinucleotide 5′ to the target site constituted an outgroup both in the *K*D and target site accessibility distributions. We also determined the extent of correlation of the flanking dinucleotide binding affinity values with a range of positions and window lengths across from the miR-1, finding a slightly greater correlation with flanking dinucleotide *K*D for a 10 nt window spanning miRNA positions 1–10 (*r2* = 0.83). Owing to the complete orthogonality of these two values (one being the averaged result of a computational algorithm applied to each read in the input library, and the other being a biochemical constant derived from enrichment of motif-containing reads in our experiments with respect to the input library) we interpret these data as evidence that the primary cause of the difference in binding affinity due to flanking sequence context is target site accessibility, as it is challenging to compose an alternative model by which the AGO–miRNA complex itself has nucleotide preferences that so accurately recapitulates the predicted accessibility profile.

To further discriminate between target site accessibility and differential Ago-miRNA–mediated primary nucleotide affinity as leading to the range of flanking dinucleotide *K*D values, we sampled 8mer-containing reads from the sequenced input RNA library such that the frequency of each of the 256 flanking dinucleotides matched that of the 72 pM AGO2–miR-1 sample, reasoning that this should recapitulate the ~11-fold increased in the geometric mean target site accessibility observed in the 72 pM AGO2–miR-1 sample library if the average predicted target site accessibilities were not causally responsible for the 100-fold range in the measured flanking dinucleotide *K*D values despite being highly correlated. This resampling explained 19.4% of the difference (Figure 3D, left). Resampling to recapitulate the flaking dinucleotide frequencies in the other four AGO2–miR-1 samples produces the same effect, with the percentage growing marginally to to 22% for the 720 pM AGO2–miR-1 sample (Figure 3D, right). By contrast, sampling 8mer reads from the input weighted directly by their relative target site accessibility values (a probability between 0 and 1), recapitulates the flanking dinucleotide frequencies (*r*2 = 0.83) (Figure S3H), Interestingly, sampling in this way produces a target site accessibility distribution far in excess of that observed within the 72 pM AGO2–miR-1 RBNS sample (Figure S3H, left). When adjusting the weighting of each read in the sampling so as to match the mean and variance of the target accessibility within the Ago-miR-1 RBNS library (see Methods), the predicted flanking dinucleotide abundances were still highly correlated with the measured abundances within the Ago-miR-1 RBNS library (*r*2 = 0.80), albeit with a compressed range (Figure S3I). Taken together, we interpret these data as demonstrating that miRNA target site sequence context has massive influence on miRNA–target binding affinity, and that the mechanistic basis of this influence is predominantly the propensity for the overall site to form secondary structure, rather than any intrinsic primary nucleotide preference by the Ago–miRNA complex itself.

**AGO-RBNS binding affinity measurements explain miRNA-specific differences in target site repression *in vivo*.**

We next sought to evaluate the utility of our AGO-RBNS derived binding affinity measurements for the prediction of miRNA–mediated repression *in vivo*. We ectopically expressed miRNA miR-1, let-7a, miR-155, miR-124, and lsy-6 in HeLa cells by transient transfection of synthetic small RNA–duplexes and performed RNA-Seq to monitor transcriptome-wide expression changes due to each of these miRNAs. While indeed numerous studies have performed similar overexpression experiments with miR-1, miR-155, miR-124, and let-7a, we opted to generate these data sets in order to maximize the similarity of all aspects of the experimental setup other than the identity of the miRNA being overexpressed. We compared, for each of the five miRNAs, the averaged log2 fold-change across all transcripts containing a single seed site (8mer, 7mer-m8, 7mer-A1, 6mer, 6mer-m8, or 6mer-A1) in its 3′ UTR, to that of the corresponding AGO-RBNS *K*D value (Figure 4A–E). We see that in all five cases the *K*D value is predictive of target site repression (*r*2 = 0.92, 0.97, 0.95, 1.00, and 0.98, *p* = 2.37 x 10–2, 3.20 x 10–3, 1.03 x 10–3, 4.50 x 10–6, and 1.21 x 10–4, for miR-1, let-7a, miR-155, miR-124, and lsy-6, repectively), tracking the miRNA-specific differences in repression across site type. In particular, we see the variability of the performance of the 7mer-A1, most extremely represented by miR-155 and miR-124, is explained by its differential binding affinity in comparison to the 8mer, for these different miRNAs. Analysis of all five miRNAs together (Figure 4F) can be explained well a single linear model (*r*2 = 0.82), and is yet improved by the omission of let-7a from this analysis, due to the presence of endogenous let-7a in HeLa cells. When looking at the additional noncanonical sites we identified, we were not able to specifically verify any individual sitetype classes, owing both to the low-number of occurences of these sites in the transcriptome due to the length of the sites, in combination with the lower predicted magnitude of their repression in comparison to the high affinity sites (Figure S4A–E). However, we still see that the relationship between binding and repression is maintained. These data suggest a model in which all miRNA-specific differences can be attributed to binding affinity, with the downstream steps in the pathway occurring as a function of occupancy of the site. They also demonstrate that the essential resource in quantitatively predicting the effects of a miRNA effects *in vivo* are binding affinity measurements for that miRNA.

**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)

**Determination of miRNA–target site *K*D values from Equilibrium AGO–RNA Bind-N-Seq**

Upon binning the read counts associated with a particular equilibrium AGO–RBNS experiment according which target site–types they contain, we determined *K*D values for each of the site types as well as for reads lacking a site type by maximum likelihood estimation (MLE). We determine values for  such that the log-likelihood of observing the data given



,

in which  is the probability density of , is maximized. Since no standard probability distribution will directly calculate the probability density of the observed count data as an explicit function of the site-type *K*D values and any additional biochemical parameters, we separate the likelihood estimation across two sequential function evaluations:



where

 .

This allows is the a the concentration of each site–type within the RNA recovered from each sample to be predicted as a function of a particular set of model parameters, and subsequently applied to a likelihood function estimating the agreement of the model predictions with the experimental data. We use a log-likelihood cost function

,

allowing MLE estimation to proceed by minimizing  with any of a number of numerical optimization routines. We now describe the functions  and  in detail.

1. The function calculates , the concentration of each site–type containing RNA  (hereafter referred to as “site”) recovered within equilibrium binding reaction sample**, where and . This function takes as input, the *K*D value corresponding to each site , the total concentration of each of the site types , the stock concentration of AGO–miRNA complex (“hereafter referred to as “AGO”), the concentration of library RNA recovered nonspecifically by nitrocellulose filter binding, , and the integer  designating which sample within the dilution series is to be simulated. Because the vector of total target–site type concentrations are fixed during the optimization routine, they are not included in the parameter vector , nor is the integer *j,* as it is specified in the subscript index. In order to allow the optimization routine to proceed unconstrained (i.e., to allow the parameter values to be anywhere in the range from –∞—+∞, the parameter vector  supplies the natural logarithm of each of the aforementioned biochemical parameters to . The recovered concentration of site  in sample  is given by



where  and are the concentration of the AGO–bound and nonspecifically bound forms of the site, respectively. By making the assumption that only the unbound sites can participate in nonspecific binding, and that the total amount of nonspecific binding is a constant ,  can be substituted, yielding



where  (100 nM), , and .

To solve for , we apply the definition of *K*D



where  and  represent the aforementioned AGO-bound and AGO-unbound forms of site  in sample , respectively, and  represents the concentration of free AGO in sample , given by the root of the following equation:



where is the total amount of AGO in the corresponding sample , given by



We note that the ostensibly arbitrary prefactor in equation accounts for the dilution series in each experiment (i.e, yields 40, 12.65, 4, 1.26, and 0.4 for  = 1, 2, 3, 4, and 5, respectively). Equation is rearranged to



­­­and substituted into equation , yielding



a function that is only in terms of the model parameters to be optimized, the experimentally derived parameter vector , its sum, and the free ago terms  *.*

2. In order to obtain MLE values of , the cost function  is required to scale monotonically with the negative log-likelihood of the site-type concentrations *x*, given theobserved counts *y* in the sequencing data for that experiment:



We remind the reader that both  and  are 2-dimensional data arrays, where  and  represent the model-predicted concentration and the integer number of read counts of site  from dilution sample  for a particular equilibrium AGO–RBNS experiment. Each column , representing one sequencing library, comprise count data emitted, whose expectation we model with a multinomial probability distribution:

**Error! Objects cannot be created from editing field codes.**

where each  is the frequency of each site, trivially given by



where  as before. The overall likelihood function  is given by the natural logarithm of the product of [2.b] for each of the paired columns of  and :





Since the third and fourth terms in equation [2.d] are unrelated to, they have no influence on the MLE estimate of . We therefore define



as the relevant cost function for our optimization routine.

3. MLE Optimization. Prior to beginning each MLE optimization routine, the parameter values of  are fixed; given by 100 nM multiplied by the frequency of each site in the input library. The parameter vector  is initialized by randomly sampling from a norma probability distribution with mean = 0, and standard deviation = 3, with the exception of the *K*D corresponding to RNAs containing no site to the miRNA, which was given a fixed value of 1 nM in all optimizations since it was determined that the model was constrained only with respect to the relative *K*D of each site relative to the no-site *K*D.

We performed the MLE estimation using the optim function in R using the L–BFGS–B method, which is a gradient based-method employing box constraints. In order to efficiently measure large numbers of *K*D values within reasonable time scales (fewer than days), as well as the instability of numerical approximation of the gradient by finite-differences, we calculated the analytical gradient of the cost function  as a function of . The derivative can be written in its most general form using the chain rule:



where  refers to the partial derivative of the cost function with respect to the model-predicted concentration of recovered site  in equilibrium sample ,  refers to the partial derivative of each model-predicted concentration with respect to the parameter vector ,  refers to the partial derivative of the concentration of each recovered site  with repect to the concentration of each bound site , and  refers to the derivative of the concentration of AGO-bound site  with respect to the parameter vector . The two partial derivative terms are found by differentiating equations



and



respectively, where is the Kronecker delta function defined by:



Substituting and rearranging yields:



The background parameter term only factors in at the level of , and not at , causing  to be zero. The derivative is therefore

 .

For the *K*D and [AGO] parameters, , and



and



since

Taking the sum of both sides of equation(?), substituting in equation (?) as well substituting in  yields



whereand substituttng   yields



hich for To solve for , we substitute  and  to equation (?), giving



and further substituted into equation (?) for the final, full equation



To solve for , we substitute in  and  for all *i*, yielding



olve for which can be substituted into equation(?) to yield:



**FIGURE LEGENDS**

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

**(A)** Purified AGO2–miR-1 is incubated with excess randomized library until binding equilibrium (left). The AGO-bound library molecules are isolated by applying the reaction to a nitrocellulose filter under vacuum (middle), and then used to prepare DNA amplicons for RNA-Seq (right). Five different concentrations of AGO2–miR-1 in the binding reaction were sequenced, in addition to the input, and a mock binding reaction with no AGO2–miR-1.

**(B)** Fractional abundance of reads containing a complementary 8mer (purple), 7mer-m8 (red), 7mer-A1 (blue), 6mer (cyan), 6mer­-m8 (light purple), or 6mer-A1 (light blue) site to miR-1, as well as reads containing none of these site types (black), in the input library and the 7.6 pM AGO2–miR-1 library. Dashed vertical lines depict the enrichment or depletion of each category of reads in the AGO-bound library compared to the input, and the dashed diagonal line represents x = y.

**(C)** (Left) Enrichment of reads containing either of the six complementary site–types to miR-1 or none as a function of protein concentration across the five AGO2–miR-1 containing samples. Colors correspond to the site type–categories described in (B). Points represent the actual enrichment values generated from the RBNS experiment, while the lines represent simulated enrichment value produced from a mathematical model of the experiment. (Right) Dissociation constant (*K*D) values obtained from the mathematical model. Reported for each measurement is the geometric mean value ± the 95% confidence interval when resampling the read data and fitting the model 200 times.

**(D)** The fraction of total AGO2–miR-1 bound library molecules corresponding to each site-type category predicted by the mathematical model, as a function of AGO2–miR-1 concentration. The five points associated with each line represent the model simulation at the five AGO2–miR-1 concentrations at which the RBNS experiments were performed. Colors are as in (B).

**(E)** (Left) Experimentally generated (points) and simulated (lines) enrichment of a list of site type–categories that has been expanded to include those identified by *de novo* identification of enriched motifs within the experiment. Site type–categories present in (B–D**)** have the same colors. The legend is ordered such that it represents that of the enrichment values.

**(F)** *K*D values obtained with the *de novo* site list and mathematical model, classifying each site 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 sequence (gray). The point and error bars corresponding to each site type–category represent the geometric mean value ± the 95% confidence interval when resampling the read data and fitting the model 200 times.

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

**(A–D)** *K*D values obtained when performing AGO–RBNS with purified AGO2–let-7a (A), –miR-155 (B), –miR-124 (C), and –lsy-6 (D). Each site type is classified by 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 (light 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 with no clear complementarity to miR-1 sequence (gray). The point and error bars corresponding to each site type–category represent the geometric mean value ± the 95% confidence interval when resampling the read data and fitting the model 200 times.

**Figure 3: Flanking dinucleotide sequence context leads to 100–fold differences in miRNA–target binding across miRNA sequence and target site types, due to binding site accessibility.**

**(A)** Experimentally generated (points) and simulated (lines) enrichment of target site–types from AGO–RBNS with miR-1, but with the 8mer–containing reads in each library further separated into 256 categories according to the identity of the 5' and 3′ flanking dinucleotide sequences. 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 G-rich flanking sequences are preferentially green. The gray point and line combinations represent the non-8mer site–type categories depicted in (1D).

**(B)** *K*D values obtained for each site type across all flanking dinucleotide sequence contexts. Each row corresponds to the *K*D values for all flanking dinucleotide sequences for a particular site type, obtained by further separating the reads corresponding to that site type and applying the mathematical analysis depicted in (A), with each point representing the geometric mean value when resampling the data and fitting the model 200 times. The points are colorized by flanking nucleotide content as in (A).

**(C)** *K*D values obtained for each of the canonical 6, 7, and 8–nt site types for miR-1, let-7a, miR-155, miR-124, and lsy-6. Each rectangle represents the one miRNA, site-type, and flanking dinucleotide combination, with its color representing the fold-difference in the *K*D value for that flanking dinucleotide context relative to the geometric mean value of that miRNA and site type. The dendrogram represents the hierarchical clustering of the 256 flanking dinucleotide context possibilities using the Pearson correlation of their log-transformed, mean-adjusted values across miRNA and site-type combination.

**(D)** Relationship between the modeled *K*D value and the average structural accessibility of the target site within corresponding library RNA molecules, across all 256 flanking dinucleotide 8mer contexts, for miR-1. The strength of the relationship is measured with the coefficient of determination given by the square of Pearson correlation coefficient *r*. The dotted line depicts the regression line of the two log-transformed quantities.

**(E)** (Left) The cumulative distribution of structural accessibility of 8mer-containing library reads in the input library (solid black line), the \_\_\_\_ nM AGO2–miR-1 library (solid red line), and in the input library when resampled to match the flanking dinucleotide frequencies within the 8mer-containing reads in the \_\_\_\_ nM AGO2–miR-1 library (dashed red line). The single point found within each line represents the geometric mean of the corresponding distribution. (Right) The expected change in structural accessibility when resampling the input library 8mer containing reads to match the flanking dinucleotide content in each of the five AGO2-miR-1 bound RNA libraries, as a fraction of the full change in structural accessibility for each respective AGO2-mIR-1 bound library compared to the input. The left-hand most bar depicts the effect change visualized with the cumulative distribution plots (left).