RNA

A PUBLICATION OF THE RNA SOCIETY

MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action

Jiayu Wen, Brian J. Parker, Anders Jacobsen, et al.

RNA published online March 9, 2011
Access the most recent version at doi:10.1261/rna.2387911

MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action

JIAYU WEN,[1,2,3] BRIAN J. PARKER,[1,3] ANDERS JACOBSEN,[1,2] and ANDERS KROGH[1]

[1]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark
[2]The Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark

ABSTRACT

Microarray expression analyses following miRNA transfection/inhibition and, more recently, Argonaute cross-linked immunoprecipitation (CLIP)-seq assays have been used to detect miRNA target sites. CLIP and expression approaches measure differing stages of miRNA functioning—initial binding of the miRNP complex and subsequent message repression. We use nonparametric predictive models to characterize a large number of known target and flanking features, utilizing miRNA transfection, HITS-CLIP, and PAR-CLIP data. In particular, we utilize the precise spatial information provided by CLIP-seq to analyze the predictive effect of target flanking features. We observe distinct target determinants between expression-based and CLIP-based data. Target flanking features such as flanking region conservation are an important AGO-binding determinant—we hypothesize that CLIP experiments have a preference for strongly bound miRNP–target interactions involving adjacent RNA-binding proteins that increase the strength of cross-linking. In contrast, seed-related features are major determinants in expression-based studies, but less so for CLIP-seq studies, and increased miRNA concentrations typical of transfection studies contribute to this difference. While there is a good overlap between miRNA targets detected by miRNA transfection and CLIP-seq, the detection of CLIP-seq targets is largely independent of the level of subsequent mRNA degradation. Also, models built using CLIP-seq data show strong predictive power between independent CLIP-seq data sets, but are not strongly predictive for expression change. Similarly, models built from expression data are not strongly predictive for CLIP-seq data sets, supporting the finding that the determinants of miRNA binding and mRNA degradation differ. Predictive models and results are available at http://servers.binf.ku.dk/antar/.

Keywords: microRNA; target determinants; microRNA transfection; CLIP-seq; HITS-CLIP; PAR-CLIP

INTRODUCTION

MicroRNAs (miRNAs) exert their post-transcriptional regulatory function primarily by destabilizing messenger RNAs (Guo et al. 2010) and also suppressing protein translation, or a combination of both mechanisms (Flynt and Lai 2008). Numerous sequence features have been proposed as important for miRNA–target interaction. In metazoa, miRNAs commonly form imperfect base pairing to 3′ untranslated regions (UTRs) of target genes as part of a complex with Argonaute protein (micro-ribonucleoprotein [miRNP]), although targets in the coding regions are also found. Often there is almost perfect base pairing to the seed of the miRNA, which is defined as bases 2–8 from the 5′ end of the mature miRNA (Lai 2002; Enright et al. 2003; Lewis et al. 2003; Doench and Sharp 2004; Rajewsky and Socci 2004). This seed pairing has been considered the key factor in miRNA−target interaction and shows strong correlation with expression changes (Lim et al. 2005; Grimson et al. 2007; Nielsen et al. 2007). Seed sites are short and are likely to occur often by chance and, therefore, computational prediction using seed-matching alone suffers from a relatively high number of false-positive predictions. The most common way to reduce this problem has been to require evolutionary conservation of target sites (Krek et al. 2005; Lewis et al. 2005; Stark et al. 2005), because conserved target-binding sites are thought more likely to be biologically functional. Additionally, the presence of multiple sites in a 3′ UTR that is targeted by one or multiple miRNAs has been shown to affect message repression (Enright et al. 2003; Doench and Sharp 2004; Grimson et al. 2007; Hon and Zhang 2007; Saetrom et al. 2007).

However, seed pairing is not the only determinant of miRNA targeting. It has been shown that the sequence context

[3]These authors contributed equally to this work.

Reprint requests to: Anders Krogh, The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark; e-mail: krogh@binf.ku.dk.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2387911.

is important for functional target sites (e.g., Didiano and Hobert 2006). MicroRNA targets have a preference to reside in AU-rich regions (Grimson et al. 2007) or AU-rich 3′ UTRs (Robins and Press 2005a). There is also evidence that miRNAs may preferentially target genes with longer 3′ UTRs (Sandberg et al. 2008) and target near the ends of 3′ UTRs (Gaidatzis et al. 2007; Grimson et al. 2007; Majoros and Ohler 2007), although this interacts with the dynamic and tissue-specific nature of UTR length due to alternative polyadenylation (Sandberg et al. 2008). Mutational studies of miRNA–target interactions show that contextual features flanking the target site can have substantial effect. For example, in a study of lys-6 miRNA in *Caenorhabditis elegans* Didiano and Hobert (2008) found that regions immediately downstream of the target site are important for enabling miRNA regulation. Moreover, this effect was independent of general flanking AU-enrichment. It is hypothesized that such sites could represent binding sites for modulating factors such as RNA-binding proteins (RBPs) (Didiano and Hobert 2008; Jacobsen et al. 2010). In addition, thermodynamic stability of the miRNA–mRNA interaction has been used to identify target sites (Enright et al. 2003; Lewis et al. 2003; John et al. 2004; Rehmsmeier et al. 2004; Krek et al. 2005). Target accessibility as measured by the free-energy cost required to open the target site has also proven useful in recognizing functional target sites (Robins and Press 2005b; Kertesz et al. 2007; Long et al. 2007; Obernosterer et al. 2008; Tafer et al. 2008).

The knowledge of miRNA target-site determinants has been largely obtained from measurements of expression changes after miRNA transfection. Lim et al. (2005) first used a miRNA transfection microarray experiment to detect the down-regulation of a large number of transcripts after overexpressing miRNAs, and other studies followed (e.g., Wang and Wang 2006; Grimson et al. 2007; Linsley et al. 2007). Similarly, miRNA knock-down followed by mRNA expression measurements has been used (e.g., Krützfeldt et al. 2005; Frankel et al. 2007). Proteomics experiments further show that for a substantial number of genes, such mRNA destabilization is highly correlated with the resultant protein repression (Vinther et al. 2006; Baek et al. 2008; Selbach et al. 2008; Guo et al. 2010).

Two recent studies (Hausser et al. 2009; Hong et al. 2009) have compared miRNA target features determined from such expression analyses with the results of Argonaute immunopurification (RIP-chip) experiments, and noted differences in the target features. For example, in contrast to earlier studies which have found that 3′ UTR length is increased in miRNA targets, these studies found that a distinguishing feature of miRNP binding was short 3′ UTR length. These studies relied upon direct Argonaute pull-down of transcripts by RNA immunopurification, which requires strongly associated RNAs and will miss loosely or transiently associated miRNPs. More recently, cross-linking with immunoprecipitation (CLIP-seq) methods have been used as an assay for miRNA target sites (Chi et al. 2009)—this method involves in vivo UV cross-linking of the mRNA and miRNP, immunoprecipitation, and isolation of cross-linked RNA segments followed by cDNA sequencing. For example, two recent CLIP-seq studies—AGO HITS-CLIP (Chi et al. 2009) and AGO PAR-CLIP (Hafner et al. 2010) methods—were shown to identify miRNA–target interactions with relatively high specificity (Chi et al. 2009, Supplemental information). It is of interest to compare the features of miRNA targets revealed by these new CLIP-seq techniques to expression-based analyses to reveal complementary feature determinants in miRNA targeting, effects of miRNA concentration, and to highlight possible selection biases in these protocols.

A limitation of previous studies is the restriction of their analyses primarily to statistical significance, with no measure of predictive performance of individual features. Although a feature may be significantly over-represented among target genes, it does not necessarily follow that it adds power to prediction of target sites, because such statistical enrichment is a function of sample size. Using predictive power to quantify miRNA–target features is informative, as it gives a direct answer to the importance of the feature and feature interactions for unseen data. For example, most features examined in this study also have highly statistically significant *P*-values for mean difference between target and nontarget sets, but their predictive power varies greatly.

In this study, we perform a comprehensive comparison of miRNA targeting features for expression-based data sets (mRNA/proteome expression following miRNA transfection/knock-down) and two CLIP-seq-based methods (HITS-CLIP and PAR-CLIP). These two types of data highlight two aspects of miRNA targeting: initially miRNAs bind to target sites through a miRNP complex, and subsequently cause message repression through degradation and/or translational changes. Accordingly, the expression-based data sets detect target genes with expression substantially regulated by miRNAs, while the CLIP-seq data sets detect target genes bound by one or more miRNPs. These two aspects of miRNA targeting may reveal, or be determined by, different miRNA–target features (Hausser et al. 2009). Hausser et al. (2009) found that target accessibility is associated with miRNP binding in RIP-chip data, while sequence composition, especially U frequency in entire transcripts, is associated with mRNA degradation. To investigate target determinants in current CLIP-seq data in comparison to miRNA transfection data, we defined and analyzed a large number of target features that may be relevant for miRNA targeting using both types of data. We used a nonparametric machine learning method to rank and analyze features by their predictive power and to investigate multivariable interactions of feature categories. Specifically, we analyzed (1) relative importance ranking of individual features, (2) dependencies and interactions of combinations of categories of features, and (3) the overall performance of predictive models trained using these features.

RESULTS AND DISCUSSION

Comparisons of predictive power reveal differences in feature importance between expression-based and AGO CLIP-seq data sets

We compared three data sets for investigating miRNA target determinants: miRNA transfection data containing mRNA expression profiles following overexpression of 12 miRNAs in HeLa cells from Lim et al. (2005) and Grimson et al. (2007), AGO HITS-CLIP data in mouse brain from Chi et al. (2009), and AGO PAR-CLIP data in HEK293 cells from Hafner et al. (2010). These experiments cover more than 50 different miRNAs in various cell lines or tissues. To examine feature predictive power, we defined positive and negative sets for all three data sets. For the expression-based data, the positive set was defined as genes with fold change above a defined threshold. For the CLIP-seq data, the positive set was defined as genes with one or more CLIP clusters for the top 20 most-abundant miRNA families (see Materials and Methods). Negative sets were randomly selected to be equally sized from genes of low fold change or no CLIP clusters present (unbound in CLIP-seq), respectively. The features were organized into six categories by their sequence, structure, or positional characteristics as listed in Table 1 and illustrated in Figure 2A, below, with details of the calculations in the Supplemental methods. The area under ROC curve (AUC) was used to measure the predictive power, and the Hanley-McNeil test was used to measure statistical significance of AUC increases unless otherwise specified (see Materials and Methods). *P*-values of significance of AUC differences of a feature between data sets are shown in Supplemental Table 2.

A practical advantage of CLIP-seq is that it provides not only transcript-level information on miRNP binding, but also localizes the target interactions on the 3′ UTR. This spatial information was utilized for the CLIP-seq data sets to improve the accuracy of estimates of flanking-feature effect sizes where applicable. For the transfection data sets and CLIP-seq negative controls where this spatial information is not available, the best match seed site was used (potential selection biases in these controls were controlled for statistically, see Materials and Methods).

We first estimated feature importance by training a predictive model (Random Forest, see Materials and Methods) and measuring the decrease in predictive power after effectively removing each feature, by randomization of the feature

TABLE 1. miRNA target and contextual features used in this study and their predictive power measured in AUC

			AUC[a]		
Category	Region	Features	miRNA transfection	HITS-CLIP	PAR-CLIP
Conservation	70 nt	Conservation of flanking region	0.65	0.80	0.76
	seed	Conservation of seed	0.70	0.80	0.79
	target	Conservation of entire target sequence	0.69	0.81	0.76
Target complementarity	target	miRNA–target sequence alignment score (John et al. 2004).	0.82	0.75	0.76
	seed	Seed type, 6-mer, 7-mer-m8, 7-mer-A1, and 8-mer (Grimson et al. 2007)	0.82	0.75	0.74
	seed	Seed mismatch/GU pairing	0.76	0.76	0.74
	out-seed	3′ out-seed pairing (Grimson et al. 2007)	0.58	0.52	0.59
3′ UTR features	3′ UTR	3′ UTR length	0.66	0.79	0.77
	3′ UTR	Number of target sites	0.78	0.75	0.73
	3′ UTR	Relative distance to 3′ UTR ends	0.56	0.61	0.61
	3′ UTR	Minimum distance to 3′ UTR ends	0.60	0.67	0.65
Composition in target site flanking regions	70 nt	Mononucleotide frequencies	0.73	0.65	0.72
	70 nt	Dinucleotide frequencies	0.71	0.69	0.69
	30 nt	AU content (Grimson et al. 2007)	0.70	0.63	0.75
	70 nt	Base compositional entropy	0.63	0.59	0.61
Target free-energy-based features	target	Free-energy loss $\Delta\Delta G$ that indicates target site accessibility (Kertesz et al. 2007)	0.72	0.68	0.67
	target	ΔG_{duplex} miRNA–target hybridization energy	0.67	0.64	0.65
Flanking strand asymmetry	70 nt	Folding free-energy difference ΔG between two strands; mean, standard deviation, and 4th moment (Wen et al. 2007)	0.69	0.57	0.56
	70 nt	Base asymmetry bias, A vs. U and G vs. C (Wen et al. 2007)	0.64	0.57	0.61
	70 nt	G+U content (Wen et al. 2007)	0.63	0.54	0.61

[a] *P*-values of significance of AUC differences of a feature between data sets are shown in Supplemental Table 2.

(mean decrease in Gini index). Figure 1A shows the overall importance ranking of various miRNA–target features (in decreasing order) for both the miRNA transfection experiments and CLIP-seq experiments. Figure 1B shows the predictive performance, measured in AUC, of individual features across each experiment. This univariate feature AUC was directly calculated from the ordering induced by each feature (see Materials and Methods).

Amongst the top-ranked features in the miRNA transfection data were, consistent with the previous literature, the seed-related features (see Table 1): miRNA-target alignment score, seed type, and number of target sites. In the expression-based data sets, the AUC heatmap (Fig. 1B) shows that the top-ranked seed-related features performed consistently across most miRNA families and spanned different types of experiments, including mRNA/protein expression following miRNA overexpression/inhibition. In contrast, both of the CLIP-seq data sets showed a different feature importance ranking, where the conservation features and 3′ UTR length appeared consistently across different miRNA families to have the strongest discriminability, while seed-related features ranked relatively lower. In both the miRNA transfection and CLIP-seq data, the highly ranked target contextual features included: target-site accessibility, flanking AU composition/U nucleotide frequency, and flanking conservation.

Conservation of target site and flanking region are the most important features in CLIP-seq targets

We examined the discriminability of sequence conservation measured by average conservation (phastCons score) for seed, entire target site, and flanking region in all three data sets. In both the CLIP-seq data sets, the three conservation regions showed strong discrimination (AUC 0.76–0.81) (Fig. 2B) and ranked higher than seed site features (see Fig. 1A). In the PAR-CLIP positive set, for example, 58% of transcripts had highly conserved seed regions with a conservation score ≥0.9 compared with only 15% in the negative set (a ratio of 3.9-fold; $P < 1\text{E-}15$, Fisher exact two-sided test). The enrichment for highly conserved CLIP clusters was also noted in Chi et al. (2009) and Hafner et al. (2010). The flanking sequence conservation also showed a marked difference in ratio of the positive set to the negative set (6.6-fold; phastCons score ≥0.9; $P < 1\text{E-}15$, Fisher exact two-sided test), as also previously observed in other studies (Nielsen et al. 2007), suggesting that the extended flanking region is important for target recognition. However, these conservation features showed only

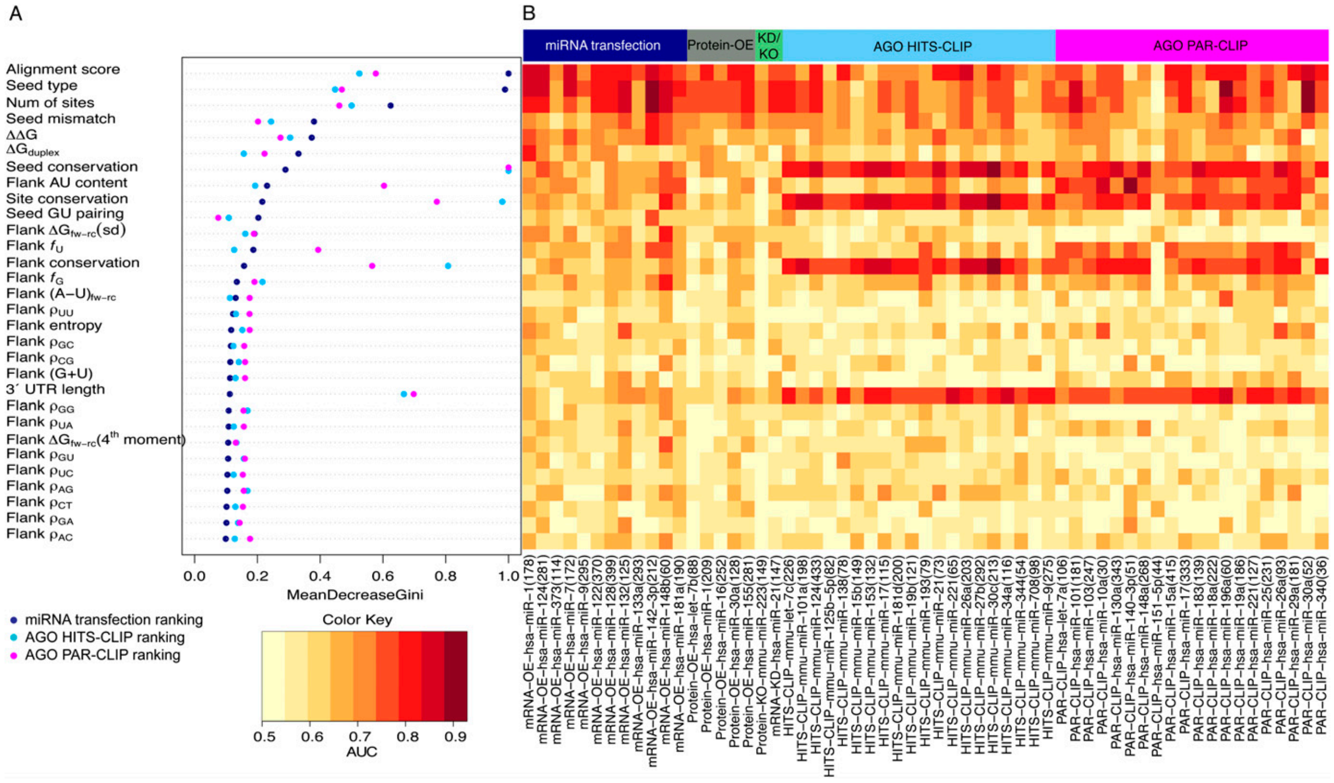

FIGURE 1. Feature importance ranking and predictive power for miRNA transfection/inhibition, HITS-CLIP, and PAR-CLIP data sets. (*A*) Feature importance rankings were evaluated using the Gini impurity criterion (normalized between 0 and 1) from Random Forest classification for three data sets, separately. Top 30 features are shown: miRNA transfection data (dark blue, in decreasing order), HITS-CLIP (light blue), and PAR-CLIP (pink). (*B*) The heatmap shows the AUCs for the corresponding features across individual data sets and miRNA families (number of 3′ UTRs in each data set in brackets). AUCs from high to low are represented by colors from red to white.

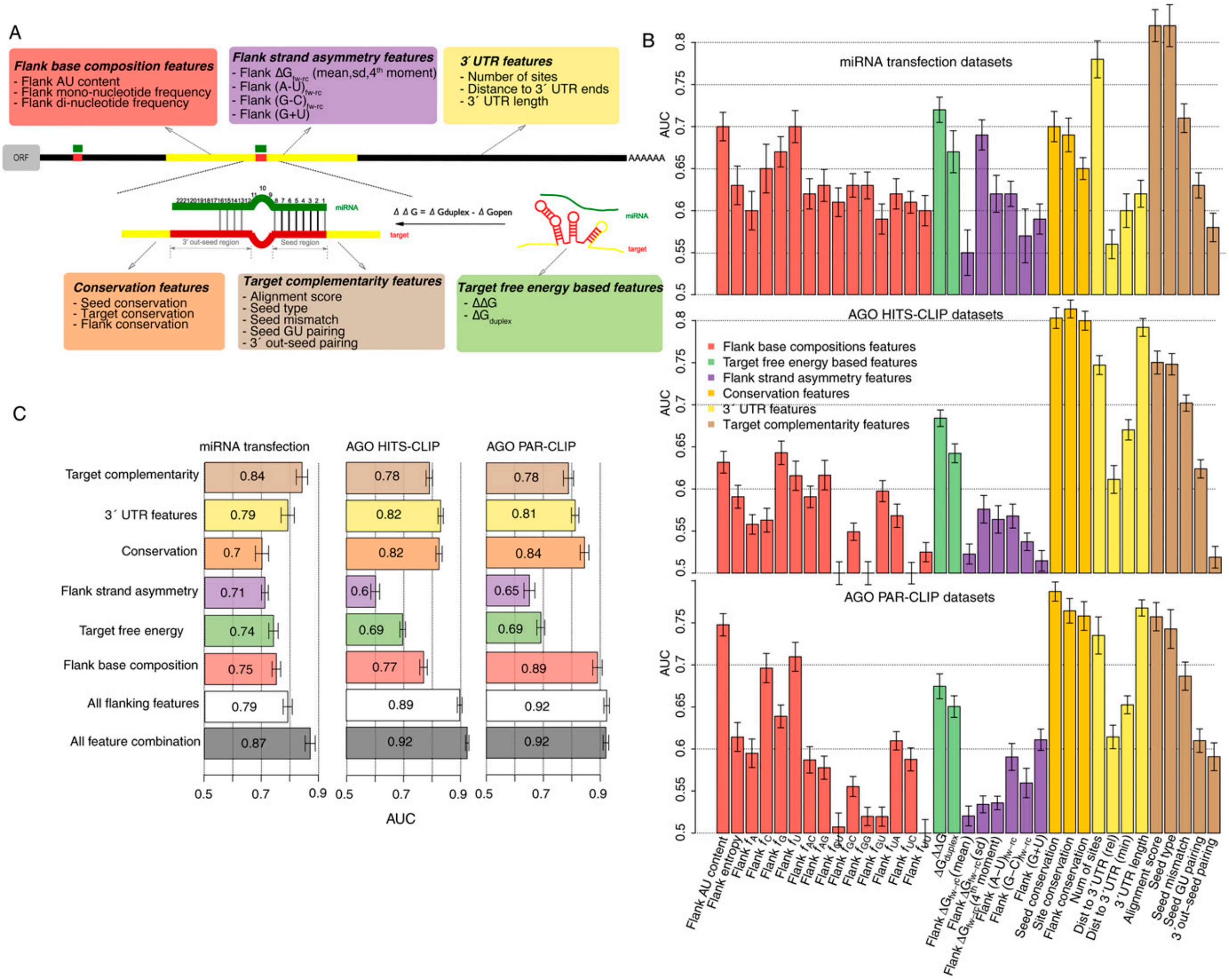

FIGURE 2. Predictive power of both individual miRNA-target features and feature combinations. (*A*) miRNA-interaction features are grouped into six categories by their sequence, structure, or positional characteristics (listed in the colored box). (*B*) Barplot shows univariate feature AUC for the positive set vs. negative sets in miRNA transfection, HITS-CLIP, and PAR-CLIP data sets. The error bars show standard errors of AUC. (*C*) Comparison of combined predictive performance for each feature category. AUCs for feature combinations in each of six categories are shown in bars for all three data sets (leave-one miRNA family-out cross-validation; Random forest classifier). Two additional feature combinations are also shown: the combination of all flanking features and the combination of all features. All flanking features include flanking nucleotide composition, target free-energy-based features, flanking strand asymmetry, flanking region conservation, and relative/minimum distance to 3′ UTR ends. Bar colors in *B* and *C* have the same color schema as *A*.

moderate discrimination, with AUC of 0.65–0.70, in the miRNA transfection data set. The conservation features increased the overall target predictive power by only 0.2% (not statistically significant; $P = 0.39$, Hanley-McNeil test) when combined with other features in the miRNA transfection data, whereas they substantially improved the predictive performance by 4.5% ($P = 3.2E-6$) in the HITS-CLIP data and by 2.2% ($P = 0.01$) in the PAR-CLIP data.

The CLIP-seq studies would be expected to preferentially detect target sites with strong miRNP interactions—these are a function of not only the miRNA component complementary to their targets but also the protein component, with Argonaute binding to the target and flanking regions, so that features such as flanking conservation become relevant. We hypothesize that, in addition, other RNA-binding proteins associated in vivo with the miRNP complex and target flanking regions would be expected to enhance this effect. RBPs can, by structural changes, make adjacent target sites accessible (Brodersen and Voinnet 2009). Other associated RBPs may dynamically inhibit or otherwise modulate the effect of the target site in vivo (e.g., Kedde et al. 2007)—we hypothesize that such target sites will be preferentially detectable by the CLIP-seq protocol, but not necessarily by single time-point transfection studies. Supplemental Figure 1 shows that binding affinity (indicated by the number of cross-linked miRNPs) is correlated with conservation.

Seed match features are more predictive in the miRNA expression-based data than AGO CLIP-seq data

In the miRNA transfection data set, a good AUC (0.82) for both the seed type and alignment score features indicates that the seed has an important role in identifying highly repressed miRNA target binding sites, which is consistent with previous miRNA transfection microarray studies. The miRNA knock-down data shows qualitatively similar results. The seed type feature showed a weaker predictive power in both the HITS-CLIP data (AUC = 0.75 vs. 0.82; P = 4.4E-7, Hanley-McNeil test) and PAR-CLIP data (AUC = 0.76 vs. 0.82; P = 9.3E-6). For example, for messages in the miRNA transfection positive set, 72% possessed at least one 7-mer seed site compared with 55% (P < 1E-15, Fisher exact two-sided test) in both the HITS-CLIP and PAR-CLIP positive sets. The miRNA-target sequence alignment score measures the quality of alignment along the target site, but heavily weighted around the seed region (John et al. 2004), giving an estimate of seed binding affinity; this feature similarly shows weaker discrimination in both CLIP-seq data sets (Table 1). The presence of multiple sites in miRNA–target interactions may simultaneously destabilize targeted messages, as observed in other studies. In the miRNA transfection data, this was confirmed by the multiple-sites feature alone yielding a good discrimination (AUC = 0.78); but, this feature showed a weaker predictive power in both the CLIP-seq data sets (AUC 0.73–0.75; P < 1.5E-5), indicating that this is primarily a determinant of expression change. Seed mismatch and GU pairing were one of the lowest ranked features by importance for both CLIP-seq data, but were higher ranked for miRNA transfection data (see Fig. 1A). This is consistent with the finding of Hong et al. (2009) of no enrichment for GU pairing or mismatches in RIP-chip data.

Transfection studies directly reveal the degradation response mechanism of particular miRNAs by manipulating the mature miRNA levels to high levels, often beyond physiological concentration, and then typically measuring the response through fold change at a single time point. Therefore, the target features highlighted are those most directly involved in mediating this degradation response. For miRNA knock-down expression studies, miRNA concentration remains at physiological levels. In contrast, CLIP-seq methods depend on the general interaction of Argonaute and potentially associated RNA-binding proteins, sufficient for cross-linking, across all endogenous miRNAs at physiological concentrations. Therefore, features determining overall miRNP binding to mRNA targets are highlighted.

To further analyze these differences between expression and CLIP-seq methods, we constructed a set of miRNA targets defined using knock-down expression data, and divided the data by fold change into 10 bins. (The knock-down expression set was defined by transcripts showing substantial up-regulation after inhibiting the top 25 miRNAs [>1.2-fold change] and having 7-mer or better seed match. See Materials and Methods.) We then plotted the overlap with PAR-CLIP–identified targets for each bin. Figure 3 shows only a small association between fold change and detection by PAR-CLIP (change in proportion of 12.5% from first to last decile by linear model fitting; P = 0.04, proportion trend test); although, overall, there is large (81%) overlap between transcripts identified by miRNA inhibition data and those identified by PAR-CLIP. This is again consistent with the miRNP-binding determinants as detected by PAR-CLIP being largely independent of those determining mRNA degradation fold change response.

The size of the effect of the stochastic interaction of miRNPs with the target mRNA would be expected to depend not only on target features of the mRNA–miRNA interaction, but also on the input concentration of miRNAs. We performed a stratified analysis of the PAR-CLIP data into high (top 10 highly expressed miRNA families) and low (bottom 10 miRNA families) miRNA expression groups, to determine the extent to which input miRNA concentrations are associated with the various features. We observed several differences between the two groups (Fig. 4A), but the most striking difference was a large change in AUC of target complementarity (AUC 0.78 vs. 0.67; P = 2.5E-9) (see Fig. 4B), indicating higher complementarity targets increasing in importance with increasing miRNA concentration. Potentially, additional sites may also be targeted at high miRNA concentration. This higher discriminability of alignment score and seed type in the group of highly expressed miRNAs is more similar to the miRNA transfection studies, whereas these features show weaker discriminability in the low miRNA expression group.

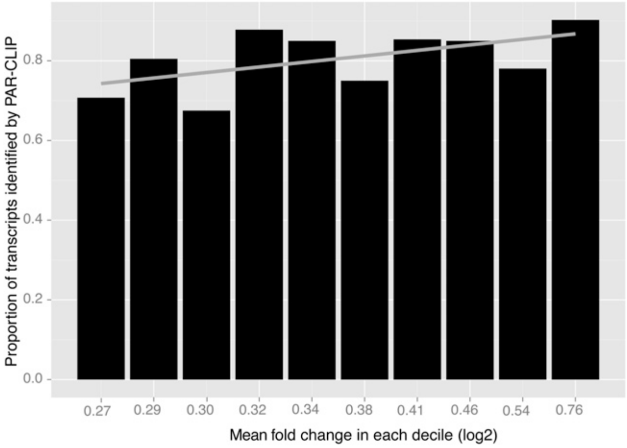

FIGURE 3. Association of targets detected by PAR-CLIP and targets detected by knock-down expression, stratified by fold change. The distribution for each decile of fold change after inhibiting the top 25 miRNAs in HEK293 cells (x-axis) versus the proportion of transcripts showing at least one PAR-CLIP cluster (y-axis) are plotted. The least-square fitted line for the proportion is shown.

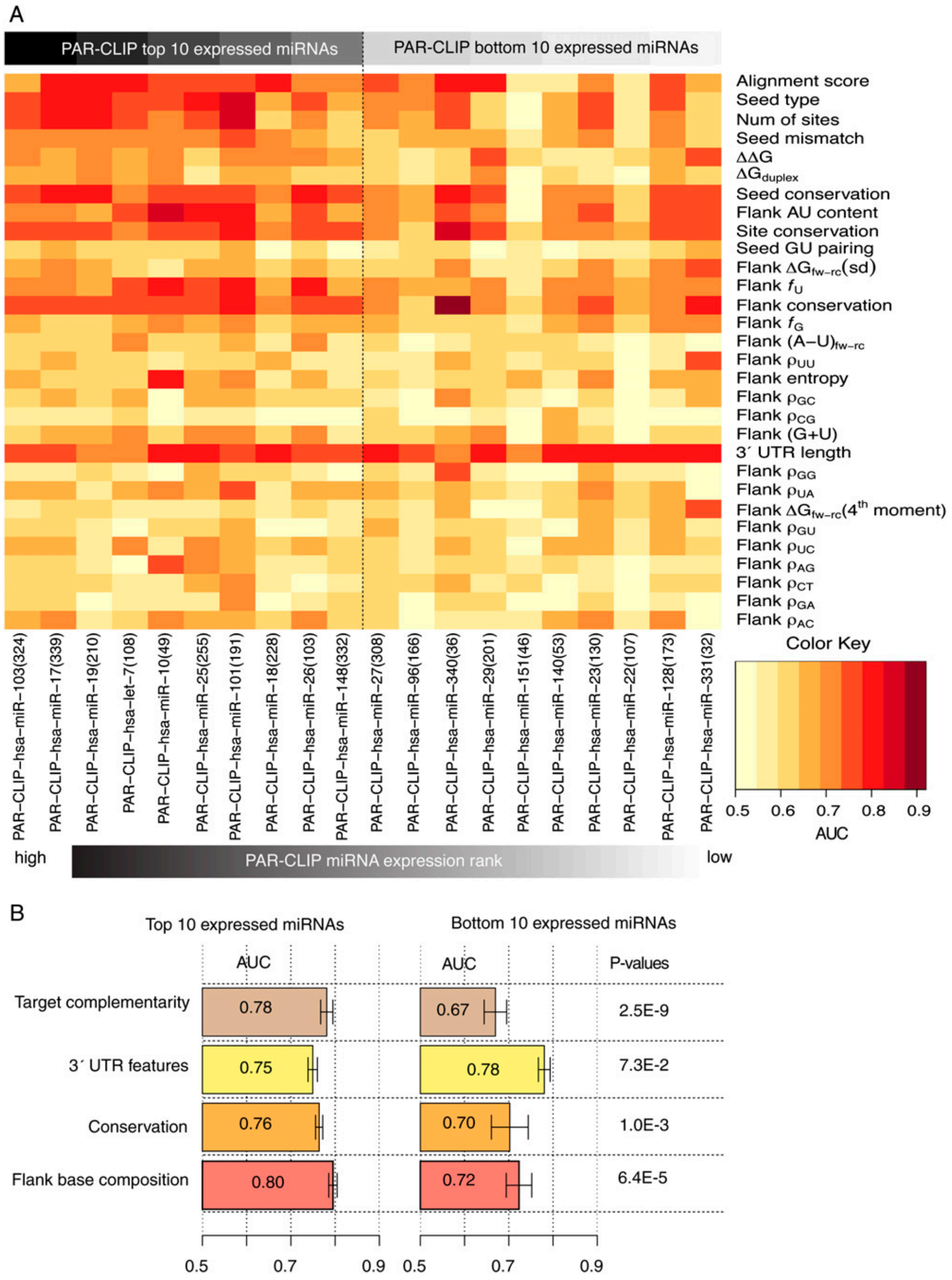

FIGURE 4. Comparison of feature predictive power between PAR-CLIP groups of the top-10 and bottom-10 expressed miRNA families. (*A*) The heatmap shows the AUCs for the corresponding features across individual data sets and miRNA families. AUCs from high to low are represented by colors from red to white. (*B*) Comparison of AUCs for each feature category between the two groups. *P*-values measure the statistical significance of AUC differences.

This result demonstrates the effect of input miRNA concentration on target determination, and suggests that the high-input miRNA concentration in the transfection studies contributes to the high AUC of seed features.

3′ UTR length has high apparent discriminability in CLIP-seq data sets

In miRNA transfection data, the average 3′ UTR length is increased in the positive set compared with that in the negative set (1979 vs. 1376 bp; $P = 7.6E-30$, Mann-Whitney-Wilcoxon test). This is consistent, in direction, with previous reports that highly expressed transcripts are evolved to possess shorter 3′ UTRs to avoid miRNA targeting (Stark et al. 2005) and targets associated with longer 3′ UTRs increasing the efficacy of target repression (Sandberg et al. 2008). This is in the opposite direction of the transfection results reported in Hausser et al. (2009), which, however, were restricted to a set of shorter 3′ UTRs with exactly one seed site, but is consistent with the direction found in the comparative genomics data sets in the same study. Due to its large variance, the 3′ UTR length feature only has poor discrimination (AUC = 0.66), which is less than the (correlated) number of sites feature (AUC = 0.78; P = 4.1E-12). The mean density of target sites per 3′ UTR was substantially increased in the miRNA transfection data (mean, 1.5 targets per 3′ UTR = 0.8 targets per kilobase) compared with the CLIP-seq data (mean, 1.3 targets per 3′ UTR = 0.6 targets per kilobase for PAR-CLIP), consistent with number-of-sites being a determinant of expression fold change.

In contrast, the discriminability of 3′ UTR length was strikingly increased in both the HITS-CLIP (AUC = 0.79 vs. 0.66 in transfection data; P = 1E-14) and PAR-CLIP data sets (AUC = 0.77 vs. 0.66 in transfection data; P = 1.2-10) and was one of the highest ranked features (Fig. 1A). The average 3′ UTR length of the positive set was much longer than that of the negative set (2495 vs. 1173 bp, P < 1E-15 for HITS-CLIP; 2745 vs. 1272 bp, P < 1E-15 for PAR-CLIP; Mann-Whitney-Wilcoxon test) (see Supplemental Fig. 2 for distributions). This was also noted in Hafner et al. (2010).

It is possible that 3′ UTR length is upwardly biased in the CLIP-seq data: theoretically, a low sensitivity of the CLIP cross-linking and downstream processing could lead disproportionately to shorter UTRs being missed (as the longer UTRs with more targets are more likely to have at least one target cross-linked and detected). To quantify this bias, we performed a simulation study using a simplified model of the CLIP-seq process (see Supplemental Fig. 3; Supplemental Methods) and showed that such an effect varies inversely with sensitivity and could lead to a bias on the order of AUC 0.1 if sensitivity is low (e.g., 58% sensitivity for low-expression transcripts in Chi et al. 2009), which is on the order of the increase in effect size of this feature in CLIP-seq compared with transfection studies. To examine this further, as CLIP-seq has been shown to be more sensitive for highly expressed genes (described in Chi et al. 2009, Supplemental "Estimates of false negative rates and sensitivity"), we stratified the PAR-CLIP data into mRNAs with high and low expression. The highly expressed genes showed a lower AUC for 3′ UTR length compared with the low expression genes (Supplemental Fig. 4), consistent with the predictions of the simulation model. The results suggest that the large AUC for this feature in CLIP-seq data may be substantially upwardly biased, and the AUC determined from the miRNA transfection studies, which would not be prone to this bias, may be a more accurate estimate of its effect on miRNA targeting.

As noted, surprisingly this result conflicts with the opposite finding for RIP-chip that was shown to select for shorter 3′ UTRs (Hausser et al. 2009; Hong et al. 2009). Hong et al. (2009) noted that structural accessibility of target sites was increased in small UTRs and hypothesized that this may account biologically for the small message bias of RIP. However, the correlation of 3′ UTR length and IP enrichment from data of another two AGO RIP-chip experiments (Hendrickson et al. 2008; Landthaler et al. 2008) show inconsistent results, with one selecting longer UTRs (r = 0.12; P = 0) and another selecting shorter UTRs (r = −0.03; P = 2.9E-02). The previous results of Stark et al. (2005) and Sandberg et al. (2008), the positive correlation in the miRNA transfection data sets, and the inconsistency among RIP data sets suggest that the analysis based on RIP data could be a result of experimental selection biases. One intuitive explanation is that the RIP protocol directly immunoprecipitates messages without fragmentation, whereas the CLIP protocol fragments the messages after cross-linking, so that small messages may be more efficiently immunoprecipitated in the RIP protocol, depending on details of the assay.

Additionally, the minimum distance to the 3′ UTR ends was substantially increased in predictive power in the CLIP data (AUC 0.65–0.67) compared with the miRNA transfection data (AUC = 0.60; $P \leq$ 1.5E-7, Hanley-McNeil test), which is consistent with the importance of this feature in determining the miRNP binding reported in Hausser et al. (2009).

Target contextual features are highly predictive in both miRNA expression-based and AGO CLIP-seq experiments

We combined all flanking features (listed in Table 1) including the nucleotide composition in flanking regions, target accessibility, flanking strand asymmetry, flanking region conservation, and 3′ UTR features (excluding 3′ UTR length and number of seed sites features). The combination of all flanking features in the miRNA transfection data gave a predictive performance (AUC nearly 0.8) comparable to target complementarity features (AUC 0.84) (Fig. 2C). For both the CLIP-seq data sets, the flanking features were also strongly predictive, with contribution largely by flanking conservation (flanks vs. seed features: 0.92 vs. 0.78, P = 1.6E-49 for PAR-CLIP; 0.89 vs. 0.78, P = 2E-25 for HITS-CLIP).

The nucleotide compositions (mono- and dinucleotide frequencies, and base compositional entropy) were computed on extended 70-bp flanking regions, and AU content was computed on immediately flanking regions (30 bp, as in Grimson et al. 2007). In the transfection data set, consistent with Grimson et al. (2007), AU content ranked relatively high for the miRNA transfection data (AUC = 0.7). The nucleotide U frequency (AUC = 0.7) alone gave a predictive power equivalent to flank AU content (AUC = 0.7). This flanking U-richness was also recently reported in Hausser et al. (2009). The flanking U frequency and AU content were substantially higher in the transfection than in HITS-CLIP data, consistent with Hausser et al. (2009).

The most striking difference between the HITS-CLIP and PAR-CLIP features was in flank U frequencies and AU content (Fig. 2B). While flank AU content and U frequency are enriched in all data sets, in PAR-CLIP flank AU content was ranked substantially higher in importance compared with HITS-CLIP (AUC 0.75 vs. 0.63; P = 2.4E-19) (see Fig. 1A).

Flank U frequency also showed a similar difference between the two CLIP data (AUC 0.71 vs. 0.62; $P = 7.8E\text{-}11$). These differences could be due to the utilization of T-to-C mutations to identify targets in the PAR-CLIP protocol, which could potentially lead to a selection bias toward U-rich sequences as discussed in Hafner et al. (2010). To test this, we first compared U frequency in PAR-CLIP clusters containing T-to-C mutations with that in PAR-CLIP AGO sequencing read data as a background (see Materials and Methods). The results showed that the average U frequency in the clusters containing T-to-C mutations is significantly higher than that in the background (36.4% vs. 30.9%, $P < 1E\text{-}15$, two-sided *t*-test). To account for potential biological variance, we also tested whether such U bias would occur in different biological experiments by showing that the average U nucleotide frequency in clusters with T-to-C mutation is significantly higher compared with the background across AGO1–AGO4 experiments ($P = 9.8E\text{-}3$, one sided *t*-test). We further performed a stratified analysis, which divided the data into 3′ UTRs containing PAR-CLIP clusters with high and low T-to-C mutation counts (>12 and <2). The AU content of the 3′ UTRs containing high T-to-C mutation clusters showed a significantly increased AUC compared with those of 3′ UTRs containing low T-to-C mutation clusters (AUC 0.73 vs. 0.69; $P = 1.6E\text{-}2$, Hanley-MacNeil test). The results are consistent with a positive selection bias toward U-rich sequences in the PAR-CLIP protocol, although a negative bias in HITS-CLIP cannot be ruled out.

The free-energy change is the difference between minimum free energies (MFEs) of miRNA-target hybridization (ΔG_{duplex}) and MFE cost needed to make the target site accessible (ΔG_{open}). AU-richness near the target site may be associated with target accessibility, because A–U base pairs are less stable than G–C. However, the contribution that the target accessibility feature makes in determining miRNA target functionality, and its dependencies, has been debated, particularly when comparing the two related features: ΔG_{duplex} (Chen et al. 2009) and AU content (Grimson et al. 2007). We tested how reliably such target structural features could discriminate targets from nontargets and the interdependence between ΔG_{duplex} and AU content. For the miRNA transfection data, the discriminative power of $\Delta\Delta G$ provided moderate support in discriminating target and nontarget sequences in the expression data (AUC = 0.72) (cf. AUC 0.76 reported in Kertesz et al. 2007 using different data sets). The MFE of miRNA–target hybridization, ΔG_{duplex}, gave a substantially weaker predictive power than $\Delta\Delta G$ (AUC = 0.67 vs. 0.72 of $\Delta\Delta G$; $P = 6.8E\text{-}3$). AU content, $\Delta\Delta G$, and ΔG_{duplex} combined gave an improved AUC of 0.78. The increase in AUC over individual features was statistically significant ($P = 0.017$), suggesting that these features are not redundant. The $\Delta\Delta G$ feature also ranked highly among the target context features in both the HITS-CLIP and PAR-CLIP data sets, although they gave a weaker AUC of 0.68 and 0.67, respectively, inconsistent with the RIP-chip results of Hausser et al. (2009).

Next, we used a feature (Wen et al. 2007) based on a structural strand asymmetry between complementary strands in RNA sequences, occurring because G–U nucleotides commonly form base pairs, but the corresponding C–A nucleotides in the complementary strand do not pair, leading to a minimum free-energy difference between the strands. This basic asymmetry signal can be used as evidence for the potential formation of strand-specific RNA structures. For example, it has also been used in the EvoFold program for strand orientation prediction (Pedersen et al. 2006). Here, we examined both mean and higher moment differences in structural strand asymmetry in 70-nt flanking regions around the seed (see Materials and Methods). We also calculated the features (G−U), (A−U), and (G+U) to estimate local compositional asymmetry across strands in these regions. The structural asymmetry feature showed larger AUC (0.69) than base compositional asymmetry (0.64; $P = 8.6E\text{-}3$), suggesting that the structural asymmetry was not simply due to compositional biases.

For the miRNA transfection data, the combination of flank base and structural strand asymmetry features led to a moderate predictive performance (AUC 0.71). This signal was consistent with potential *cis*-regulatory structures being on the same strand and functional at the RNA level. This feature gave lower results in the PAR-CLIP and HITS-CLIP data (AUC 0.65 vs. 0.6; $P = 6.3E\text{-}4$).

Comparisons of target-predictive models trained on miRNA expression-based and AGO CLIP-seq data sets imply different underlying feature determinants

To measure and compare how well the multivariable combination of all of the above features predicts miRNA-target binding, we built a combined probabilistic predictive model using a Random Forest classifier. A Random Forest classifier was used, as it is known to perform well on mixed variable types and heterogeneous data and can reveal nonlinear feature interactions. We trained predictive models on the miRNA transfection, HITS-CLIP, and PAR-CLIP data sets separately (Fig. 5A; Supplemental Fig. 5). A leave-one miRNA family-out cross-validation was used in estimating the predictive performance of each model (see Materials and Methods). For the miRNA transfection data set, when the positive class was defined by an expression fold change of >1.6-fold down-regulation, most of the sequences were classified correctly with an AUC of 0.87 for identifying miRNA targets. For comparison, we also varied the thresholds for defining down-regulation as the response variable, ranging from 1.3-fold to 1.8-fold (see Supplemental Fig. 6). As expected, setting lower thresholds for down-regulation lessened the classification performance, because more noisy genes with relatively low fold change were included in the positive class. The results of both the CLIP-seq data showed a very good predictive performance for identifying miRNA targets (both AUC = 0.92). Removing the 3′ UTR length

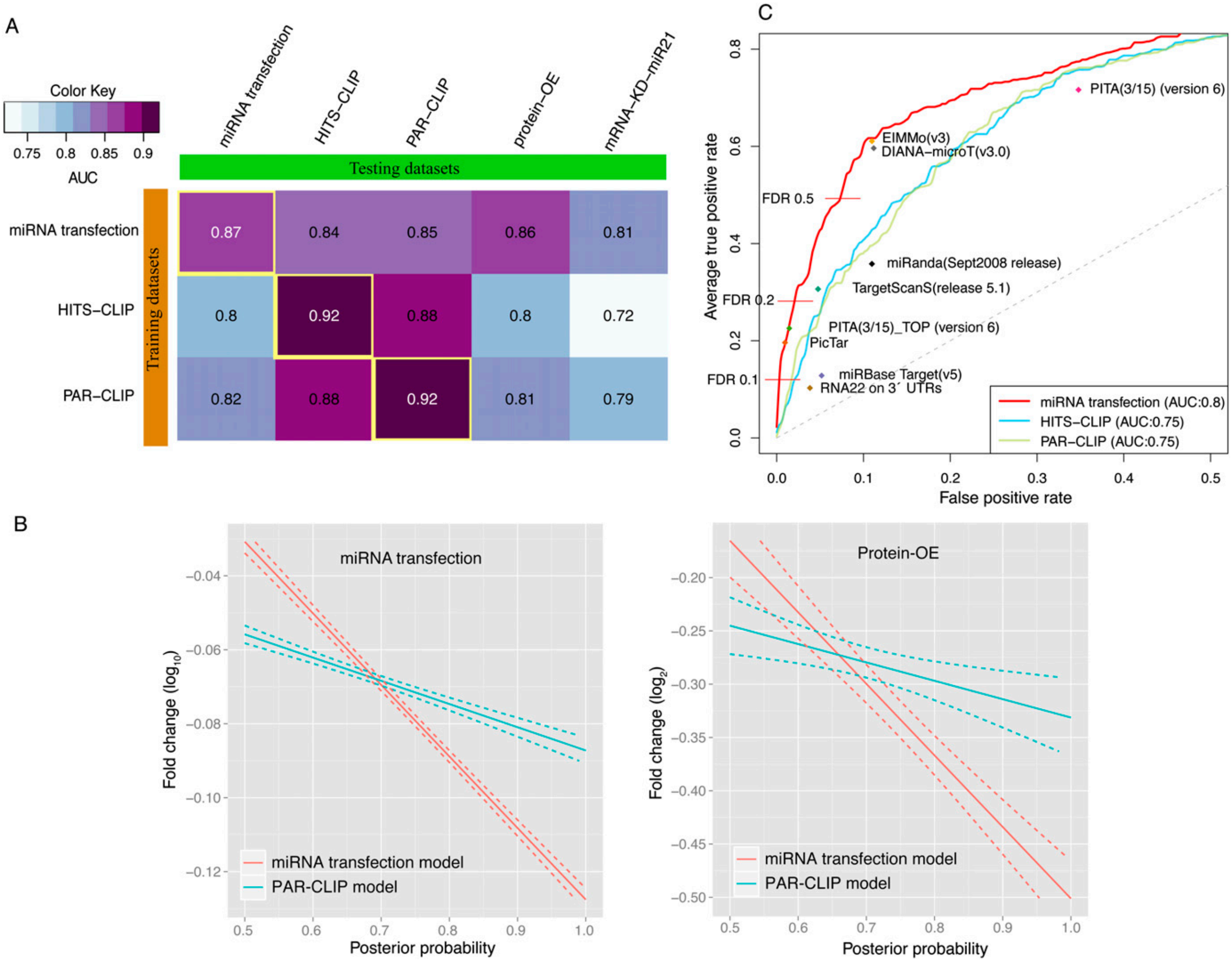

FIGURE 5. Comparisons of predictive models trained on miRNA transfection, HITS-CLIP, and PAR-CLIP data sets. (*A*) The heatmap shows the predictive performance (measured in AUC) of models using all features, trained and validated on different data sets. Diagonal cells (yellow framed) show the AUCs for leave-one miRNA family-out cross-validation of the models. Off-diagonal cells show the model trained on the row data set and tested on the column data set. (*B*) Comparisons of the correlation of expression change and posterior probability of prediction generated by the miRNA transfection-trained model (red) and the PAR-CLIP-trained model (blue). The two models were applied to miRNA transfection and protein-OE data sets without thresholding on fold change. Linear regression lines were fitted to all predicted targets with posterior probability ≥0.5 and dash lines are the 95% confidence interval of the fitting. (*C*) The performance of the model trained on the miRNA transfection data using all features and, for comparison, several current target prediction programs. All were evaluated on the protein-OE data set. The ROC curves (red) show the true positive rate (sensitivity) vs. false positive rate (1−specificity) for the positive set (note: fold change threshold ≥1.4-fold) vs. the negative set (low fold change) classifications. Short red horizontal lines on ROC curves marked predictions with a FDR limit of 50%, 20%, and 10% (FDR estimation in the Supplemental Materials and Methods), showing the trade-off between sensitivity and specificity. Colored dots represent maximum sensitivity (percentage of miRNA-target interactions predicted from the positive set) vs. 1−specificity (percentage of miRNA–target interactions predicted from the negative set) of predicted miRNA–target interactions for each corresponding prediction program. The CLIP-seq models evaluated on the same data are also shown.

feature, which is possibly affected by experimental biases (see above), did not affect the overall performance of the CLIP-seq models. Also, varying the thresholds for defining the CLIP-seq-positive class did not substantially change the predictive performance (Supplemental Fig. 6).

As an independent validation of the models, we tested whether each model trained on different data sets was capable of detecting miRNA–target interactions on test sets that were not used in the training. For each model trained on these three data sets, we compared its predictive performance against each other as test sets, as well as on two additional test sets: the pooled proteomics data for five miRNAs that measured protein changes after overexpressing miRNAs (protein-OE) and mRNA profiling of knock-down microarray experiments for hsa-miR-21 (mRNA-KD-miR21). The results (Fig. 5A) showed that the model trained on the transfection data set performed very well on the protein-OE data giving an AUC of 0.86, which is almost identical to the cross-validation results. However, both the CLIP-seq models gave a lower AUC (0.8–0.81; $P < 3\text{E-}5$). On the miRNA-KD-miR21 data, the model trained on transfection data gave an AUC of 0.81, whereas on this data the HITS-CLIP and PAR-CLIP models again gave a lower AUC of 0.72 ($P = 2.3\text{E-}10$) and 0.79 ($P = 0.12$), respectively.

Note that if the weaker CLIP-seq model predictive performance on expression data was simply due to inaccurate CLIP training data, it would lead to a weak model for predicting CLIP-seq data in cross-validation as well; however, the results do not suggest this: In cross-validation the CLIP models can predict very well (both AUC = 0.92), and the PAR-CLIP model has good predictive function for HITS-CLIP data and vice versa (both AUC = 0.88) (Fig. 5A). The finding that the CLIP-seq models showed a lower prediction performance on both expression-based proteome and miRNA-inhibition data than the miRNA transfection model provides support for the notion that the underlying determinants revealed by CLIP-seq data, i.e., for miRNP–target binding, differ from those revealed by miRNA expression-based data, i.e., degree of degradation. A predictive model trained on a RIP-chip data set (Hendrickson et al. 2008) showed qualitatively similar results to the CLIP-seq models, but with substantially lower performance across all data sets (see Supplemental Table 1), suggesting a higher specificity for the CLIP-seq protocols.

As posterior probabilities generated from the models provide an estimate of the confidence in the predictions, we further tested how well the combination of all features could infer the degree of target repression (see Fig. 5B). When predicting both mRNA and protein expression fold change, the model trained on transfection data showed a clear linear relationship between posterior probability and fold change, whereas the model trained on the PAR-CLIP data showed only a weaker association with expression change at the mRNA level (slope = −0.19, $P < 1\text{E-}15$ vs. slope = −0.06, $P < 1\text{E-}15$, respectively) and the protein level (slope = −0.67, $P < 1\text{E-}15$ vs. slope = −0.17, $P = 2.6\text{E-}3$, respectively). This adds further support for the finding that determinants of AGO binding and mRNA down-regulation vary. This is also consistent with previous reports showing that seed match correlates with fold change (Nielsen et al. 2007) and consistent with a function of miRNA as continuous modulators tuning expression rather than purely binary on-off switching (Bartel and Chen 2004; Bartel 2009).

As a point of reference, we compared the prediction of the model built from miRNA transfection data to existing target prediction programs, using protein-OE as the validation data set. The results (Fig. 5C) showed that the full predictive model with the combination of all features performed well at all false-positive levels, demonstrating that the features extracted and models used in this analysis were comparable to existing approaches and published analyses.

CONCLUSION

With the increasing use of new CLIP-seq protocols, it is of importance to understand the determinants of microRNA targeting revealed in comparison with existing techniques, both to gain a deeper understanding of the features in a wider context and to be aware of possible selection biases in the protocol. We presented a nonparametric predictive model-based analysis to investigate miRNA–target interaction determinants by comparing miRNA transfection microarray data and recent CLIP-seq data. This study gives an in-depth evaluation of miRNA–target feature importance and interaction based on both statistical significance and also, importantly, on predictive performance.

The analysis revealed clear differences in target conservation and seed match features between miRNA expression-based and AGO CLIP-seq data sets. These two types of data sets are complementary: The AGO CLIP-seq data indicates whether miRNP binds to the 3′ UTRs, whereas the expression analysis reveals features leading to a substantial expression effect size of this binding. Conservation features were shown to be highly ranked and strongly predictive for the CLIP-seq data, while seed match features, including seed type, alignment score, number of sites, GU pairing, and mismatches are most predictive in the miRNA expression-based data, due in part to higher input miRNA concentrations.

We hypothesize that the CLIP-seq protocol will preferentially select for the most strongly bound miRNP–target interactions due to the interaction of protein components of the miRNP complex with flanking regions, and this could be enhanced by additional RNA-binding protein components. This could explain the substantial conservation in the flanking regions.

The CLIP-seq analyses also showed a large increase in 3′ UTR length for the positive set. Although this ranked as one of the most important features, a selection bias toward long 3′ UTRs may account for some of this effect, so caution is needed in its interpretation. The CLIP-seq data showed substantial differences to previous RIP protocol analyses, most strikingly in an inverse effect for 3′ UTR length, probably reflecting different selection biases of the protocols. We note that the HITS-CLIP and PAR-CLIP data showed consistent results across most of the features analyzed except for U/C base compositional differences, adding support to these findings. The predictive models trained on the CLIP-seq and expression data sets showed considerable differences in predictive performance when applied to other independent proteome and miRNA knock-down expression data sets, further demonstrating the differences in the target determinants.

The combination of all flanking features was shown to be highly predictive, giving a prediction performance comparable to the seed match features. There are several lines of evidence supporting that flanking regions surrounding target sites are functionally related to miRNA targeting:

nucleotide composition, especially U-frequency enrichment in target flanking regions; conserved regions extending beyond the seed site to at least 70 bp; the structural strand asymmetry in flanking regions, due to potential G-U pairings extending to 70 bp or more.

Utilizing CLIP cross-linking protocols with a predictive analysis has provided additional evidence on the features and their interactions determining miRNP binding and subsequent message degradation: in some cases providing additional support for previously published observations, and in other cases new insights into the relative importance of target and flanking features and the performance of the CLIP-seq protocols. Predictive models and results are available at http://servers.binf.ku.dk/antar/.

MATERIALS AND METHODS

Data resources

miRNA transfection data sets

MiRNA transfection microarray expression data sets in Lim et al. (2005) and Grimson et al. (2007) were retrieved from the GEO database (accession nos. GDS1858 and GSE8501). The data sets contain mRNA expression profiles generated by overexpressing 12 miRNA duplexes in HeLa cells relative to negative controls using Agilent microarrays. Note that the miRNAs used were all from different microRNA families, and so, are expected to have low-sequence similarity. We also obtained a miRNA knock-down microarray expression data set generated by transfecting LNA hsa-miR-21 in MCF-7 cells (Frankel et al. 2007). We applied a nonspecific filtering step to exclude those genes showing low overall expression levels in the control samples, as these genes were unlikely to clearly show down- or up-regulation after miRNA transfection or inhibition. To do this, for each miRNA transfection microarray set, only those probes with an expression level over at least half of the control samples greater than the median expression level of the control samples were retained for further analysis. For the miR-21 knock-down microarray data set, we required the interquartile range of probe expression levels to be greater than the median value of the interquartile range of expression levels for all probes. Expression values were averaged for those genes with corresponding multiple probes. We obtained the pSILAC proteomics data set, which measured protein changes following transfection of five miRNAs. We also downloaded proteome data with mir-223 gene knock-out in mouse neutrophils compared with wild-type mouse from Baek et al. (2008). We used fold changes obtained from these proteome data sets directly without further data processing.

CLIP-seq data sets

Two recent publications by Chi et al. (2009) and Hafner et al. (2010) reported cross-linking Argonaute protein–RNA complex experiments, followed by immunoprecipitation in mouse brain (AGO HITS-CLIP) and in human embryonic kidney (HEK) 293 cell lines (PAR-CLIP). We downloaded the AGO HITS-CLIP Ago ternary map table (mm9) for the 20 most-abundant miRNAs from http://ago.rockefeller.edu/index.php. The PAR-CLIP data for AGO-bound clusters (AGO 1–4) was obtained from the Supplemental tables of the published study. We subsequently annotated the gene symbols of the PAR-CLIP clusters based on their genomic coordinates. We restricted the feature analysis to 3′ UTR regions for both expression and CLIP-seq data sets. Annotated 3′ UTR sequences were obtained from the UCSC table browser. We used the miRNA annotations of clusters from Chi et al. (2009) and Hafner et al. (2010). We obtained the top 100 expressed miRNA rankings in HEK293 cells from PAR-CLIP original manuscript's Supplemental Table S7 and clustered the miRNAs into 50 miRNA families, where mature sequences sharing the same 6-mer seed sites (miRNA position 2 to 7) or sharing the same family name in miRBase were grouped into the same family. From those, a representative member for each of the top 20 most-abundant miRNA families was used for this study.

PAR-CLIP AGO1–AGO4 sequencing data

To calculate PAR-CLIP background nucleotide frequencies, we used PAR-CLIP AGO1–AGO4 sequencing data (GEO: GSM545212, GSM545213, GSM545214, GSM545215) after mapping to the genome. We extracted sequence reads (added 10-bp flanks on both sides) that mapped to 3′ UTRs from all AGO data and calculated sequence nucleotide frequencies.

PAR-CLIP mRNA expression data

We calculated PAR-CLIP transcripts expression change using the top 25 miRNA inhibition against mock-transfected microarrays in HEK293 cells (GEO: GSM538818, GSM538819, GSM538820, GSM538821). Arrays were normalized (using the vsnrma package in Bioconductor). We also used mock-transfected arrays to estimate mRNA expression levels by averaging over two mock transfection replicates.

RIP-chip data

We obtained RIP-chip data in HEK293 cells (Hendrickson et al. 2008) from the original manuscript's Supplementary Tables. Similarly to the CLIP-seq data, we used the top 20 most-abundant miRNA families in the model training.

Construction of positive and negative sets

In the miRNA transfection data sets, we included only down-regulated genes for miRNA transfection and up-regulated genes for miRNA inhibition experiments. Although the true status of which genes are miRNA targets is not directly available, genes that show a change in expression after overexpressing or inhibiting a miRNA are enriched for target genes. For each miRNA transfection experiment, the positive set was defined as genes with an expression fold change beyond a defined threshold, and the negative set was defined as genes with no substantial fold change and, subsequently, constructed by randomly selecting a set of genes of equal size to the positive set. The positive set and the negative set were pooled together from all miRNA transfection experiments. We compared different thresholds of expression fold change that were used to determine the positive target set. The thresholds of fold change were set at a range from 1.2-fold to 1.8-fold down-regulation (in steps of 0.1) for the pooled miRNA transfection data sets, 1.2-fold to 1.7-fold down-regulation (in steps of 0.1) for the pooled pSILAC proteomics data set, and 1.1-fold to 1.3-fold up-regulation (in steps of 0.05) for the miR-21 knock-down mRNA data set. For both the HITS-CLIP and PAR-CLIP data, the positive set was defined as a set of genes with one or more CLIP-seq clusters present, and the negative set was constructed by randomly selecting a set of genes of equal size as the target set, but in which no CLIP-seq clusters (including clusters in UTRs and coding regions) could be found. Transcripts with cluster labeling were obtained from the data resource described above. The positive set and the negative set were balanced in size for each miRNA and pooled together from all miRNAs for each CLIP data. We applied different thresholds for defining the positive set: for the HITS-CLIP data the number of clusters was in a range of from two to 10 and a BC value (a reproducibility measure in five biological replicates) was of a range of from two to five; for the PAR-CLIP data, the number of T-to-C mutations (a measure of cross-linking efficiency) was set at a range of from greater than one to five. Three data sets were constructed for all feature analyses in this study unless otherwise specified: (1) a subset of mRNA expression data set following 12 miRNA transfection with the positive set defined by fold change $\geq$1.6-fold, (2) a subset of HITS-CLIP data set with the positive set defined by the number of tags $\geq$9 and BC $\geq$4, and (3) a subset of PAR-CLIP data set with the positive set defined by T-to-C mutations $\geq$5 for the top 20 most-abundant miRNA families. For RIP-chip data, the positive set was defined as genes with an IP enrichment of greater than fourfold, which was selected to give a data set size approximately equal to PAR-CLIP data. The negative set was defined as genes that were not bound by AGO and then constructed by randomly selecting a set of genes of equal size as the positive set.

Extraction of miRNA–target interaction features

All 3′ UTRs were extracted from the UCSC table browser and the unique longest 3′ UTR for each gene was used. We gene symbols as the gene identifiers in all experiments in this study to avoid gene identifier conversion bias (Ritchie et al. 2009).

CLIP-seq methods provide spatial information with the estimated target locations. For the miRNA transfection data sets, which do not have such spatial information available, putative target sites were predicted using miRanda (John et al. 2004) with loose cut-offs (score $\geq$70, energy ≤ -4, and default settings for other parameters). 3′ UTRs that did not pass this initial scan were excluded. For each 3′ UTR, the best target site was used for the analysis (except that the total number of target sites was also used as a feature), i.e., the target site with the highest miRNA-target alignment score. Note that this approach allows the best putative match without introducing a selection bias on seed match features, as it uses precisely the same method on both the positive and negative sets (see Supplemental Materials and Methods for miRNA–target features used in this study).

For CLIP-seq positive data sets, the spatial information of CLIP-seq cluster sites was utilized to compute all features where applicable (except for seed-related features, see below) to give the best estimates of AUC. For the CLIP-seq negative sets, where spatial information is not available, we used the best scoring target site as described above. However, selection biases could potentially be induced by different selection methods being applied between positive and negative sets for seed-related features. The results show that seed-related features are much less predictive when using CLIP-seq cluster sites to define the positive set (e.g., PAR-CLIP clusters AUC = 0.71 vs. best scoring approach AUC = 0.78; P = 5.4E-9), as the null negative sets have an upward selection bias induced from the best scoring selection method; the measured AUCs for these seed-features (target complementarity features, number of target sites, and target free-energy-based features) are biased lower (measured AUCs are shown in Supplemental Fig. 8). Therefore, the seed-related features were calculated using the best scoring seed match in both positive and negative sets to ensure that the features do not suffer from selection bias. For global features such as UTR length, these issues do not apply. To validate that the conclusions for comparisons between miRNA transfection and CLIP-seq data are not affected by these potential biases, a completely unbiased feature comparison where the best putative match sites were used in all data sets is presented in the Supplemental Material (Supplemental Figs. 7–9; Supplemental Table 3).

miRNA–target feature importance and prediction using Random Forest

To accurately quantify miRNA–target interaction features, we measured the discriminative power of both individual and combined features with receiver operating characteristic (ROC) curves. ROC curves have the advantage of showing the performance over a full range of classification thresholds. The area under the ROC curve (AUC) is a summary measure of the discriminative power of the given features for a given classifier, which varies from 0.5 for nondistinguishable classes to 1.0 for perfectly separable classes (as a guideline, a value of 0.9–1.0 indicates excellent discrimination, 0.8–0.9 indicates very good discrimination, and 0.5–0.6 indicates no useful discrimination). The Hanley-McNeil test (Hanley and McNeil 1982), which uses an estimate of the standard error of AUC for a given sample size, was used to measure statistical significance of AUC increases. A Random Forest classifier (Breiman 2001) was used to combine features. Univariate feature AUCs were directly computed from the rank orderings of each feature by formula: $\mathrm{AUC} = \frac{\sum_{i=1}^{n_1} R_{1i} - n_1(n_1+1)/2}{n_1 n_2}$, where n_1 and n_2 are sample size in positive and negative sets, respectively, and R_{1i} is the ranking of a feature in the positive set (Hanley and McNeil 1982). We calculated the univariate feature AUCs using colAUC from the caTools package in R.

We built a predictive model using a Random Forest classifier with multivariable combinations of the features. A leave-one-miRNA family-out cross-validation was used to evaluate the classifier performance. To do this, the data set was split into subsets, with a subset representing each miRNA family. By leaving one subset out in each round, a Random Forest classifier was trained on the remaining data and evaluated on this withheld subset. To evaluate the discriminability of the classes, the ROC curves and the associated AUC were calculated. To check for any biases in the machine learning evaluation, a permutation test was performed wherein the class label of high or low fold changes were randomly shuffled. It showed the expected random predictive accuracy of ~50%. We used the mean decrease in Gini index in a Random Forest classifier to estimate feature importance. The parameter settings of the Random Forest classifier were set to their defaults. For comparison, the data was also classified using a simple linear classifier—linear discriminant analysis (LDA). It performed comparably with an overall performance of model evaluation using LDA decreased by an AUC of 1%–2% (data not shown).

Additional details and methods are presented in the Supplemental Material.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank an anonymous reviewer for suggesting the stratified analysis based on miRNA concentration. We thank Kristoffer Smed for designing the web page. This work was supported by the Novo Nordisk Foundation and the Danish National Science Foundation. B.J.P. was funded by a Statistics Network Fellowship, Department of Mathematical Sciences, University of Copenhagen.

Received July 24, 2010; accepted January 20, 2011.

REFERENCES

Baek D, Vilén J, Shin C, Camargo FD, Gygi SP, Bartel DP. 2008. The impact of microRNAs on protein output. *Nature* **455:** 64–71.

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

Bartel DP, Chen CZ. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat Rev Genet* **5:** 396–400.

Breiman L. 2001. Random forests. *Mach Learn* **45:** 5–32.

Brodersen P, Voinnet O. 2009. Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol* **10:** 141–148.

Chen K, Maaskola J, Siegal ML, Rajewsky N. 2009. Reexamining microRNA site accessibility in *Drosophila*: a population genomics study. *PLoS ONE* **4:** e5681. doi: 10.1371/journal.pone.0005681.

Chi SW, Zang JB, Mele A, Darnell RB. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460:** 479–486.

Didiano D, Hobert O. 2006. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat Struct Mol Biol* **13:** 754–755.

Didiano D, Hobert O. 2008. Molecular architecture of a miRNA-regulated 3′ UTR. *RNA* **14:** 1297–1317.

Doench J, Sharp P. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev* **18:** 504–511.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in *Drosophila*. *Genome Biol* **5:** R1. doi: 10.1186/gb-2003-5-7-r1.

Flynt AS, Lai EC. 2008. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat Rev Genet* **9:** 831–842.

Frankel LB, Christoffersen NR, Jacobsen A, Lindow M, Krogh A, Lund AH. 2007. Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *J Biol Chem* **283:** 1026–1033.

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

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27:** 91–105.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466:** 835–840.

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

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve. *Radiology* **143:** 29–36.

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

Hendrickson DG, Hogan DJ, Herschlag D, Ferrell JE, Brown PO. 2008. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PLoS ONE* **3:** e2126. doi: 10.1371/journal.pone.0002126.

Hon LS, Zhang Z. 2007. The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. *Genome Biol* **8:** R166. doi: 10.1186/gb-2007-8-8-r166.

Hong X, Hammell M, Ambros V, Cohen SM. 2009. Immunopurification of Ago1 miRNPs selects for a distinct class of microRNA targets. *Proc Natl Acad Sci* **106:** 15085–15090.

Jacobsen A, Wen J, Marks DS, Krogh A. 2010. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Res* **20:** 1010–1019.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. 2004. Human microRNA targets. *PLoS Biol* **2:** e363. doi: 10.1371/journal.pbio.0020363.

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

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

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus K, Stoffel M, et al. 2005. Combinatorial microRNA target predictions. *Nat Genet* **37:** 495–500.

Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M. 2005. Silencing of microRNAs in vivo with ‘antagomirs’. *Nature* **438:** 685–689.

Lai EC. 2002. MicroRNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30:** 363–364.

Landthaler M, Gaidatzis D, Rothballer A, Chen PY, Soll SJ, Dinic L, Ojo T, Hafner M, Zavolan M, Tuschl T, et al. 2008. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA* **14:** 2580–2596.

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

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

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773.

Linsley PS, Schelter J, Burchard J, Kibukawa M, Martin MM, Bartz SR, Johnson JM, Cummins JM, Raymond CK, Dai H, et al. 2007. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol* **27:** 2240–2252.

Long D, Lee R, Williams P, Chan C, Ambros V, Ding Y. 2007. Potent effect of target structure on microRNA function. *Nat Struct Mol Biol* **14:** 287–294.

Majoros WH, Ohler U. 2007. Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics* **8:** 152. doi: 10.1186-1471-2164-8-152.

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

Obernosterer G, Tafer H, Martinez J. 2008. Target site effects in the RNA interference and microRNA pathways. *Biochem Soc Trans* **36:** 1216–1219.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2:** e33. doi: 10.1371/journal.pcbi.0020033.

Rajewsky N, Socci ND. 2004. Computational identification of microRNA targets. *Dev Biol* **267:** 529–535.

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. 2004. Fast and effective prediction of microRNA/target duplexes. *RNA* **10:** 1507–1517.

Ritchie W, Flamant S, Rasko JE. 2009. Predicting microRNA targets and functions: traps for the unwary. *Nat Methods* **6:** 397–398.

Robins H, Press W. 2005a. Human microRNAs target a functionally distinct population of genes with AT-rich 3′ UTRs. *Proc Natl Acad Sci* **102:** 15557–15562.

Robins H, Press W. 2005b. Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci* **102:** 4006–4009.

Saetrom P, Heale BS, Snøve OJ, Aagaard L, Alluin J, Rossi JJ. 2007. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res* **35:** 2333–2342.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320:** 1643–1647.

Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. 2008. Widespread changes in protein synthesis induced by microRNAs. *Nature* **455:** 58–63.

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. 2005. Animal microRNAs confer robustness to gene expression and have a signiflcant impact on 3′ UTR evolution. *Cell* **123:** 1133–1146.

Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, Hofacker IL. 2008. The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol* **26:** 578–583.

Vinther J, Hedegaard MM, Gardner PP, Andersen JS, Arctander P. 2006. Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. *Nucleic Acids Res* **34:** e107. doi: 10.1093/nar/gk1590.

Wang X, Wang X. 2006. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res* **34:** 1646–1652.

Wen J, Parker BJ, Weiller GF. 2007. In silico identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula. *In Silico Biol* **7:** 485–505.