Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C–miRNA complexes and the degradation of miRNA targets

Jean Hausser, Markus Landthaler, Lukasz Jaskiewicz, et al.

Genome Res. 2009 19: 2009-2020 originally published online September 18, 2009
Access the most recent version at doi:10.1101/gr.091181.109

Letter

Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C–miRNA complexes and the degradation of miRNA targets

Jean Hausser,[1,3] Markus Landthaler,[2,3,4] Lukasz Jaskiewicz,[1] Dimos Gaidatzis,[1,5] and Mihaela Zavolan[1,6]

[1]*Biozentrum, University of Basel and Swiss Institute of Bioinformatics, CH-4056 Basel, Switzerland;* [2]*Howard Hughes Medical Institute, Laboratory for RNA Biology, The Rockefeller University, New York, New York 10021, USA*

How miRNAs recognize their target sites is a puzzle that many experimental and computational studies aimed to solve. Several features, such as perfect pairing of the miRNA seed, additional pairing in the 3′ region of the miRNA, relative position in the 3′ UTR, and the A/U content of the environment of the putative site, have been found to be relevant. Here we have used a large number of previously published data sets to assess the power that various sequence and structure features have in distinguishing between putative sites that do and those that do not appear to be functional. We found that although different data sets give widely different answers when it comes to ranking the relative importance of these features, the sites inferred from most transcriptomics experiments, as well as from comparative genomics, appear similar at this level. This suggests that miRNA target sites have been selected in evolution on their ability to trigger mRNA degradation. To understand at what step in the miRNA-induced response individual features play a role, we transfected human HEK293 cells with miRNAs and analyzed the association of Argonaute/EIF2C–miRNA complexes with target mRNAs and the degradation of these messages. We found that structural features of the target site are only important for Argonaute/EIF2C binding, while sequence features such as the A/U content of the 3′ UTR are important for mRNA degradation.

[Supplemental material is available online at http://www.genome.org. The expression data from this study have been submitted to Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession no. GSE14537.]

Since the prediction of animal miRNA targets was first tackled computationally (Lewis et al. 2003; Stark et al. 2003), many approaches, taking into account features ranging from evolutionary conservation to the position of the putative target site and the nucleotide composition of its environment, have been proposed. The constraints that functional miRNA target sites obey, as well as the mechanism of miRNA action, are intensely debated. The initial paradigm that emerged from the study of *Caenorhabditis elegans* miRNAs lin-4 (Wightman et al. 1993) and let-7 (Reinhart et al. 2000) was that miRNAs induce translational repression. More recent studies challenged this paradigm and demonstrated that substantial miRNA-induced mRNA degradation occurs under both overexpression (Lim et al. 2005), as well as under physiological conditions (Bagga et al. 2005). This opened the possibility to study the determinants of miRNA targeting based on transcriptome-wide measurements of mRNA changes in response to overexpression (Grimson et al. 2007; Karginov et al. 2007; Linsley et al. 2007; Baek et al. 2008; Landthaler et al. 2008; Selbach et al. 2008), knockdown (Krützfeldt et al. 2005), and knockout (Zhao et al. 2007) of miRNAs. But because the ultimate readout of the miRNA activity is the protein output of the target transcripts, the natural expectation is that measurements of protein expression changes will generate the most appropriate data sets for studying principles of miRNA-target site recognition. Such data became available very recently, when Selbach et al. (2008) and Baek et al. (2008) determined the changes that are induced in the proteome profiles upon miRNA overexpression and depletion by different SILAC (stable isotope labeling with amino acids in cell culture) approaches.

Extensive previous work revealed that 7–8 nucleotides (nt) at the 5′ end of the miRNA are very important for target recognition (Lai 2002; Lewis et al. 2003, 2005; Doench and Sharp 2004; Brennecke et al. 2005). Aside from this, the sequence composition of the 3′ UTRs (Robins and Press, 2005) or of the immediate environment of the putative target sites (Grimson et al. 2007), the position of the site in the 3′ UTR (Gaidatzis et al. 2007; Grimson et al. 2007; Majoros and Ohler 2007), the base-pairing pattern in the 3′ region of the miRNA (Grimson et al. 2007), the structural accessibility of the target site (Robins et al. 2005; Kertesz et al. 2007; Long et al. 2007; Tafer et al. 2008), and the presence of multiple target sites in close proximity (Enright et al. 2003; Grimson et al. 2007) have also been reported to be predictive for the functionality of miRNA target sites. The relative importance of these features and in particular the relative contribution of sequence versus structural determinants are, at this point, intensely debated.

[3]These authors contributed equally to this work.
Present addresses: [4]Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin-Buch, Robert-Rössle-Strasse 10, 13125 Berlin-Buch, Germany; [5]Friedrich-Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH-4058 Basel, Switzerland.
[6]Corresponding author.
E-mail mihaela.zavolan@unibas.ch; fax 41-61-267-1584.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.091181.109. Freely available online through the *Genome Research* Open Access option.

In an attempt to identify the features that most generally characterize miRNA targets, we performed a systematic analysis of a number of large-scale publicly available data sets, each typically involving multiple miRNAs. The experiments, reported by Krützfeldt et al. (2005), Linsley et al. (2007), Grimson et al. (2007), Karginov et al. (2007), Selbach et al. (2008), and Baek et al. (2008), covered a variety of conditions, from miRNA overexpression to miRNA knockdown, in cell lines that expressed a normal amount of DICER1, as well as in DICER1 hypomorphs. The effects of miRNAs in these experiments were measured either at the level of the transcriptome or of the proteome. In order to clarify the steps at which different features appear to come into play, we have supplemented these published data sets with our own data on transcriptome-wide changes and Argonaute/EIF2C-bound mRNAs in miRNA-transfected cells. Finally, to better understand the nature of the selection pressure on miRNA target sites, we analyzed the same set of features for sites that we previously predicted with high and low probabilities to be under evolutionary selection (Gaidatzis et al. 2007).

Results

Characterization of target sites inferred in individual studies

The approach was to select, from each experiment, a set of functional and a set of nonfunctional sites, and to perform two-sample *t*-tests for the difference of the mean values of various features as described in the Methods section. Some of the features that we wanted to compute depend on the immediate sequence environment of the miRNA target site and we therefore only considered cases in which the mRNA-level response could be attributed with reasonable accuracy to a particular miRNA target site, for which the environment-dependent features could be unambiguously computed. Based on previous results (Lewis et al. 2005;Gaidatzis et al. 2007; Grimson et al. 2007), we thus selected the transcripts containing precisely one putative target site that matched nucleotides 1–8, 2–8, or 1–7 of the miRNA that was manipulated in the experiment. Because we found that the effect of the 3′ UTR sites was to be more reproducible (Supplemental Fig. 1) compared with that of CDS sites, we further selected those transcripts in which the putative target site was located in the 3′ UTR. Finally, we only considered sites that were located at least 100 nt away from the 3′ UTR boundaries in order to be able to compute the environment-dependent features. The results are shown in Figure 1 (see also Supplemental Fig. 2) in which each feature that we computed is a row and each individual experiment is a column. The matrix cells indicate how well individual features perform in distinguishing functional from nonfunctional putative target sites in a particular experiment. Red and blue matrix cells denote positive and negative *t*-values, respectively, i.e., cases in which the feature takes significantly higher (red) or lower (blue) values in functional miRNA target sites compared to nonfunctional target sites. For instance, the right-most column of the Figure 1 summarizes the comparison of putative miR-17 sites that have a high inferred probability with those that have a low inferred probability of being under evolutionary selection (Gaidatzis et al. 2007). The third cell from the top of that column, labeled "target site Eopen," is dark blue, meaning that the energy required to open the secondary structure of the putative target site is significantly smaller for sites with high probability, relative to sites with low probability, of being under evolutionary selection. This in turn suggests that evolutionary selection favored miR-17-complementary sites that are more accessible at the level of mRNA secondary structure. The second cell from the top of this column, labeled "target site Einteraction" is also dark blue, indicating that the energy of interaction between the miRNA and the putative target site is significantly lower (i.e., the interaction is more stable) for sites with high, relative to low, probability of being under evolutionary selection. In contrast, the third cell from the bottom of the column, labeled "flanks U content" is dark red. This indicates that the frequency of U nucleotides is significantly higher in the regions flanking the sites with high, relative to low, probability of being under evolutionary selection.

Applying two-dimensional hierarchical clustering with Ward linking on the Euclidean space of feature *t*-values reveals that the target sites inferred from most transcriptomics and from the comparative genomics data sets have similar properties. They reside in A- and U-rich sequence environments, the miRNA target region and its flanks are structurally accessible, and the binding free energy between the seed region of the miRNA and the mRNA is low. This indicates that the evolutionarily selected miRNA target sites support an mRNA degradation response to miRNAs. Strikingly, the proteomics data sets form an entirely different cluster, together with a few of the associated transcriptomics and the EIF2C (Argonaute) immunoprecipitation data sets. For this cluster the above-mentioned features are largely uninformative. This is very surprising because the targets that were identified, based on proteomics measurements, are enriched in miRNA seed matches, just as the targets that were previously identified based on transcriptomics measurements (Selbach et al. 2008).

One possible explanation for the less significant *t*-values obtained in the analysis of proteomics data sets is that the number of proteins that are sampled in the proteomics experiments is considerably lower (by a factor of 5–6) compared to the number of transcripts whose expression is measured in a microarray experiment. By scaling down the transcriptomics data sets through resampling such that we analyze similar numbers of genes from transcriptomics and proteomics experiments we found that this simple explanation does not hold (Supplemental Fig. 3). On the other hand, we found that although functional sites, identified in these experiments have, as expected, a higher probability of being under evolutionary selection compared to nonfunctional sites, the difference is less pronounced compared to that inferred from other types of experiments. This is shown in Figure 1 (feature labeled "ElMMo") for all the miRNAs covered by the proteomics experiments, and in Supplemental Figure 4 for the specific case of miRNAs that have been studied by multiple groups using a number of different technologies. This result is not due to the ElMMo algorithm having a poor ability to quantify specifically the functionality of the target sites determined through proteomics measurements, because as shown in Supplemental Figure 5, the accuracy of ElMMo in predicting proteomics data is similar to that of TargetScanS.

Although the features of functional target sites are consistent across most of the studied miRNAs, a few experimental data sets exhibit a striking reversal of the sign of the base content features, with the G and C base contents correlating positively and A and U contents negatively with site functionality. These data sets correspond to let-7 and miR-30a transfections, but not to the let-7 sites predicted based on evolutionary conservation, whose profile is consistent with that of most transcriptomics experiments. We conjecture that these observations are due to both let-7 and miR-30a inhibiting components of the RNAi pathway.

A number of studies already reported on the negative feedback that let-7 may exert on the RNAi pathway through targeting DICER1 (Forman et al. 2008; Tokumaru et al. 2008) and Selbach

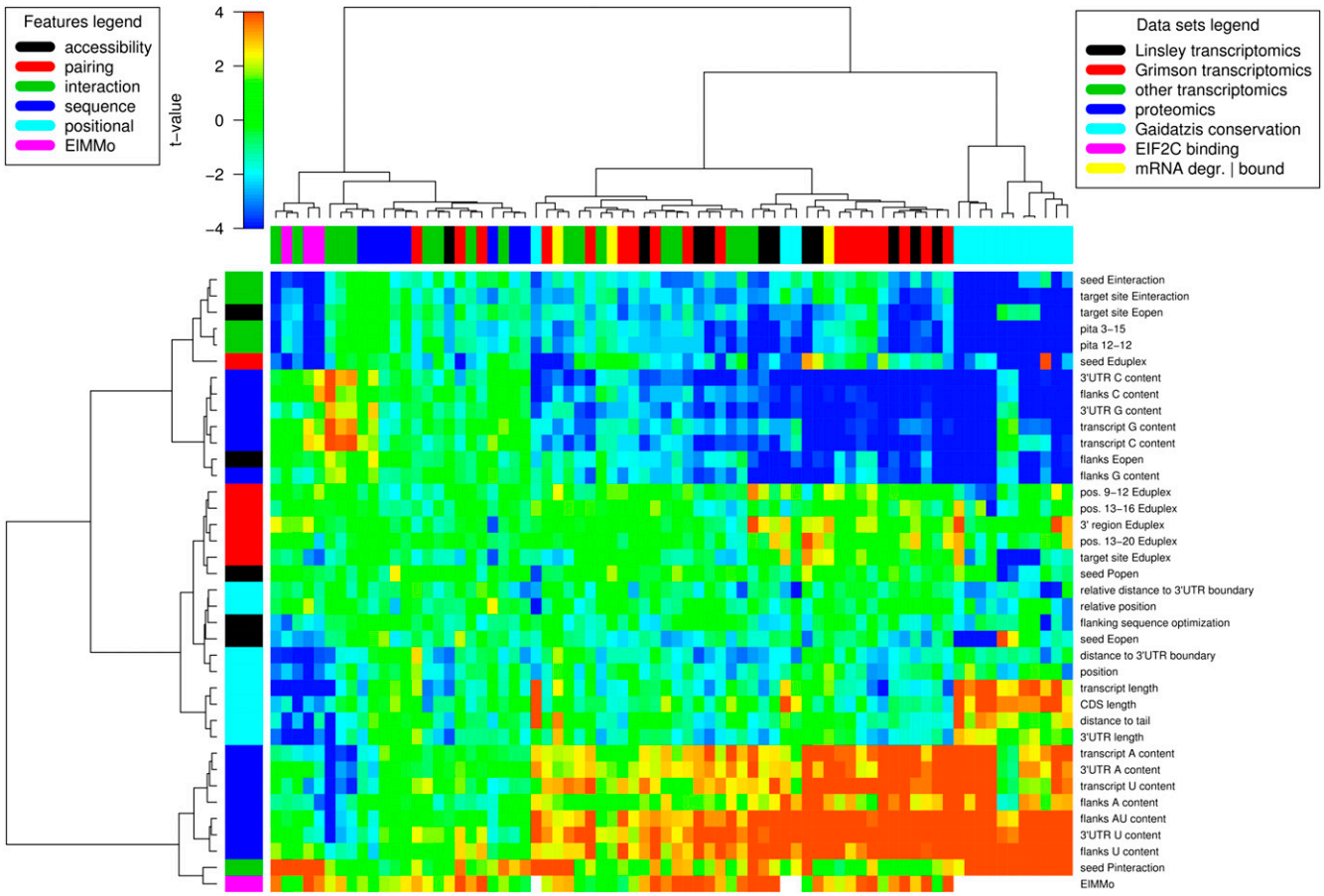

Figure 1. Predictive power of different features of putative miRNA target sites (rows) in predicting functional sites across the 74 data sets (columns). The data sets covered transcriptomics and proteomics measurements after miRNA transfection, transcriptomics measurements after miRNA knockdown, profiling of mRNAs bound to EIF2C/miRNA complexes, and target prediction based on comparative genomics. The heat-map shows the t-values comparing the distributions of feature values in functional vs nonfunctional miRNA target sites. The red color indicates positive predictors of miRNA functionality, while the blue color negative predictors of miRNA functionality. The dendrograms of features and data sets were produced through hierarchical clustering using Ward linkage on the Euclidean space of *t*-values. See also Supplemental Figure 2, in which the data set represented in each column is also indicated.

et al. (2008) already demonstrated that the DICER1 protein level increases strongly (over fourfold) upon let-7 knockdown. Similarly, we suggest that miR-30a targets the P-body component and EIF2C interactor TNRC6A (also known as GW182), which carries four matches to the miR-30a seed in its 3′ UTR, all of which are conserved all the way from human to chicken, and whose mRNA level decreases by 21% upon overexpressing miR-30a (Selbach et al. 2008). The consequence of overexpressing these miRNAs may therefore be to antagonize the effects of endogenously expressed miRNAs. Thus, the transcripts that are identified as let-7 and miR-30a targets by virtue of their down-regulation in the transfection experiments are probably transcripts that do not carry functional seed matches to miRNAs endogenously expressed in the cell, which would otherwise result in their increased expression in response to a general down-regulation of the RNAi pathway. If this were the case, we would expect the transcripts that are down-regulated in let-7 and miR-30a transfections to be depleted in functional target sites for the endogenous miRNAs. This is precisely what we found when we analyzed the transcriptomics data from Selbach et al. (2008): the transcripts that are down-regulated in the let-7 and miR-30a transfections are significantly depleted of evolutionarily selected sites for the miRNAs that are abundantly expressed in HeLa cells (Supplemental Fig. 6). One may argue that a similar behavior would be produced by the saturation/competition effect recently described by Khan et al. (2009). This effect however would apply to all the transfected miRNAs, not only to the let-7 and miR-30a, which is not what we found. We rather observed that in the miR-1, miR-155, and miR-16 transfection experiments, the mRNAs that were most down-regulated following the miRNA transfection were enriched in evolutionarily selected sites for the abundant HeLa miRNAs. The competition between the transfected and the endogenous miRNAs still occurs generally across all transfection experiments, but it is only observable at the earliest time points (Supplemental Fig. 7).

To further establish that TNRC6A is a target of miR-30a, we cloned the TNRC6A 3′UTR into a luciferase vector and we transfected this into HeLa cells with or without simultaneously transfecting the miR-30a antisense inhibitor. Transfection of the TNRC6A reporter results in a significant reduction (48%) of the luciferase activity compared with the transfection of an empty vector (Supplemental Fig. 8), while simultaneous transfection of the miR-30a antisense inhibitor results in almost complete relief of repression. This result supports our initial conjecture that the reversal of the sign of the sequence features in the miR-30a transfection experiment

is due to the negative feedback that miR-30a exerts on the miRNA pathway.

The experiments that measured the binding of EIF2C2 protein (also known as Ago2) to mRNAs resulted in the second category of data sets that exhibited the reversal of sign for the sequence features. In these data sets, the G and C contents of the transcripts and of the miRNA target site environment also correlated positively with site functionality (in this case EIF2C2 binding), while the A and U contents were negative predictors. Compared with the let-7 and miR-30a transfections, these correlations were however weaker and not significant statistically. On the other hand, structure features such as the accessibility of the miRNA binding site and the energy of interaction between the miRNA and mRNA were good predictors of the functionality of these sites, as they were for the sites inferred from transcriptomics or comparative genomics analyses. The next section describes our detailed investigation of the features that favor EIF2C2 binding and those that favor subsequent mRNA degradation.

Because the sequence composition of the environment of the miRNA target site affects the structural accessibility of the site, it is currently unclear which of these features is primarily undergoing evolutionary optimization. To address this question, we shuffled the sequence flanking functional miRNA target sites (keeping the miRNA target site fixed) and we asked whether the energy required to open the structure of the miRNA target site was higher in the context of the shuffled sequences compared to the real sequence. We found that for a subset of the data sets this is indeed the case (see Fig. 1, "flanking sequence optimization"), providing weak but statistically significant support to the hypothesis that the sequence surrounding functional miRNA target sites is constrained to increase the accessibility of the miRNA target site beyond what can be explained from the A, C, G, U content of the flanking regions (see also Fig. 2). The fact that this property does not generally characterize all data sets explains, in part, the current controversies concerning the relative importance of sequence and structure parameters in determining the functionality of miRNA target sites (Grimson et al. 2007). We further found that with the exception of the accessibility of the miRNA flanking regions, which correlates with the G + C content of these regions, the sequence features that we computed do not correlate well with the structure features (Supplemental Fig. 9). This and the results in the next section suggest that sequence and structure features come into play in a nonredundant manner, at different steps of the RNAi effector cascade, and that it is probably necessary to take them both into account in order to understand miRNA targeting specificity.

Finally, comparative genomics-based analyses reported that miRNAs tend to target transcripts with long 3′ UTRs (Stark et al. 2005). Strikingly, here we found that functional miRNA target sites that are identified experimentally generally reside in transcripts with relatively short 3′ UTRs, and that the transcript length is an even better predictor of functionality compared to the 3′ UTR length. Nonetheless, within the long 3′ UTRs in which evolutionarily selected sites are found, functional sites reside closer to the 3′ UTR boundaries (stop codon or poly-A tail) compared to nonfunctional sites, as has been previously reported (Gaidatzis et al. 2007; Grimson et al. 2007; Majoros and Ohler 2007).

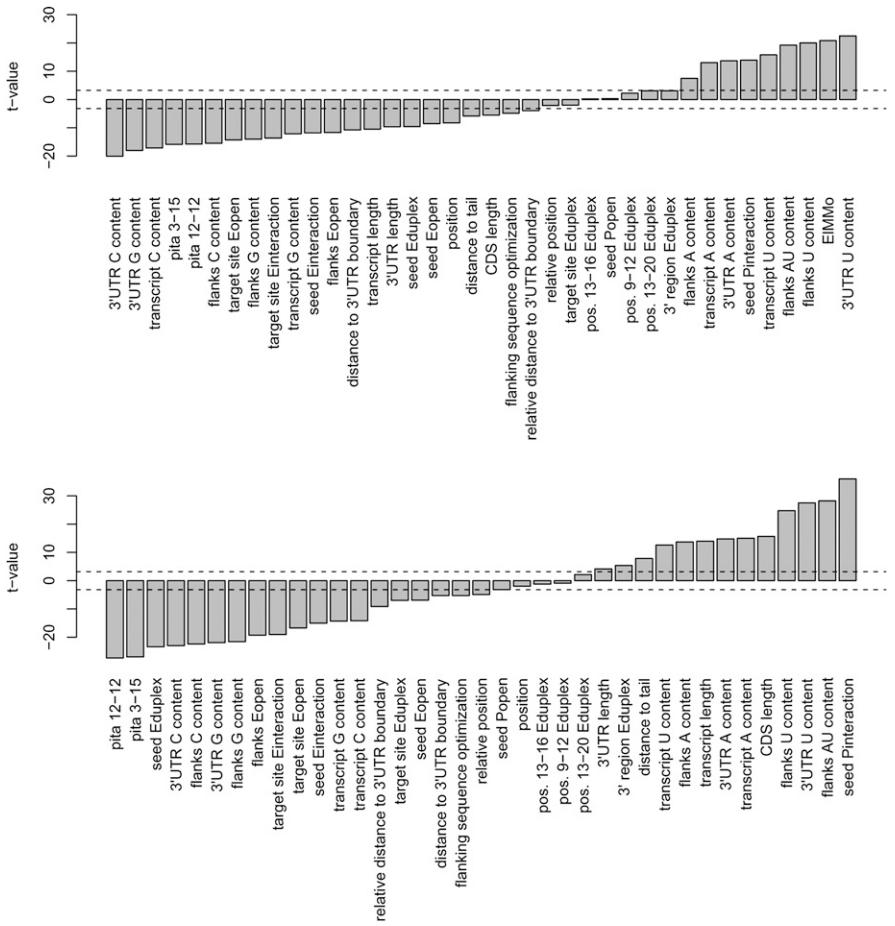

Figure 2. (*Upper* panel) Predictive power of different features across all transcriptomics experiments, excluding the let-7 and miR-30a transfections. (*Lower* panel) Predictive power of different features across all comparative genomic data sets. The *y*-axes show the *t*-values of the individual features when comparing their distribution in functional vs nonfunctional sites, aggregating over all data sets. The dotted horizontal lines represent the cutoff, where the *t*-values are significant with a bilateral type I error of 5% after applying the Bonferroni multiple testing correction. Pita 12–12 and Pita 3–15 are the scores according to the algorithm described in Kertesz et al. (2007), using 12–12 or 3–15 nt upstream and downstream of the miRNA target site for computing target site accessibility.

Structural features direct EIF2C2 binding while sequence features are associated with mRNA degradation

To gain insight into the origin of the sequence and structure biases discussed above, we transfected HEK293 cells stably expressing EIF2C2 with either a miRNA (miR-124 and miR-7) or a mock control, and we measured the mRNA expression in total RNA and in the RNA from the EIF2C2 immunoprecipitate (IP) with oligonucleotide microarrays. The degree of miRNA-specific EIF2C2 association and degradation of individual mRNAs were quantified

by the enrichment of the respective mRNAs in the EIF2C2 immunoprecipitates and the total cellular RNA, respectively, of miRNA-transfected compared to mock-transfected cells (see Supplemental Material). We also analyzed the results of a similar experiment performed with miR-124 by Karginov et al. (2007).

Binding to EIF2C2 of transcripts whose 3′ UTRs contain precisely one match to the miRNA seed was very reproducible between the two biological replicates of each transfected miRNA, with correlation coefficients of 0.85 for miR-124 and 0.70 for miR-7 (Fig. 3, upper panels). Moreover, the degree of EIF2C2 binding was correlated with that of mRNA degradation (Fig. 3, lower panels, $r = -0.70$ for miR-124, $r = -0.62$ for miR-7), with the large majority of EIF2C2-bound transcripts undergoing some degree of degradation. This correlation between EIF2C2 binding and mRNA degradation was much higher than the correlations that were reported between changes in the mRNA and in the protein levels by Selbach et al. (2008) and Baek et al. (2008) (see also Supplemental Fig. 10). A small number of EIF2C2-bound mRNAs did not show evidence of degradation, as previously reported by Karginov et al. (2007) and Hendrickson et al. (2008). To experimentally confirm that such transcripts are nonetheless regulated by miR-124 and miR-7, we generated dual luciferase reporter constructs containing the 3′ UTRs of some of the EIF2C2 bound mRNAs. Cotransfections of these luciferase reporters with the respective miRNA resulted in a reduction of luciferase activity compared to control transfections, indicating that irrespective of whether they undergo degradation, EIF2C2-bound transcripts are translationally repressed by miRNAs (Supplemental Fig. 11).

We returned to the features that we tested on the targets inferred from all other experiments and asked at what step, mRNA binding or degradation of bound mRNAs, do these features come into play. As shown in Figure 4, the *t*-statistics for the energy necessary to unwind the secondary structure of the seed pairing region (labeled "seed Eopen") and of the entire target site (Fig. 4, "target site Eopen") were significantly negative, meaning that they were significantly smaller in EIF2C2 bound sites compared with unbound sites. That is, we found that 3′ UTRs that are specifically bound by EIF2C2 tend to have seed- and miRNA-binding regions that are structurally more accessible, consistent with the results previously reported by Ameres et al. (2007). The energy of hybridizing the seed (Fig. 4, "seed Eduplex") makes a major contribution to EIF2C2 binding. Combining the structural accessibility of the seed-binding region with the energy of hybridizing the seed to the target site into a probability of interaction gives the most significant difference between target sites that are and those that are not bound by the EIF2C2-containing RISC complex. Note that we found the same feature to be highly predictive of miRNA sites that are under evolutionary selection (Fig. 1). The 3′ region of the miRNA, on the other hand, does not appear to play a crucial role in EIF2C2 binding to miRNA seed-complementary sites (Fig. 1, "pos. 13-16 Eduplex"), consistent with the whole miRNA hybridization energy (Fig. 1, "target site Eduplex") being a weaker determinant of EIF2C2 binding than the energy of hybridizing the seed (Fig. 1, "seed Eduplex"). None of these features however, was able to distinguish between *bound sites* that do and those that do not promote degradation.

In contrast, we found that features describing the sequence composition of the transcripts harboring miRNA target sites have a dramatic effect on the degradation of bound transcripts. While at the level of EIF2C2 binding, the nucleotide composition does not appear to play a statistically significant role, once transcripts are bound by EIF2C2, it is the U, and to a smaller extent the A, content that is a positive predictor of mRNA degradation (Fig. 4, lower panels). The trends

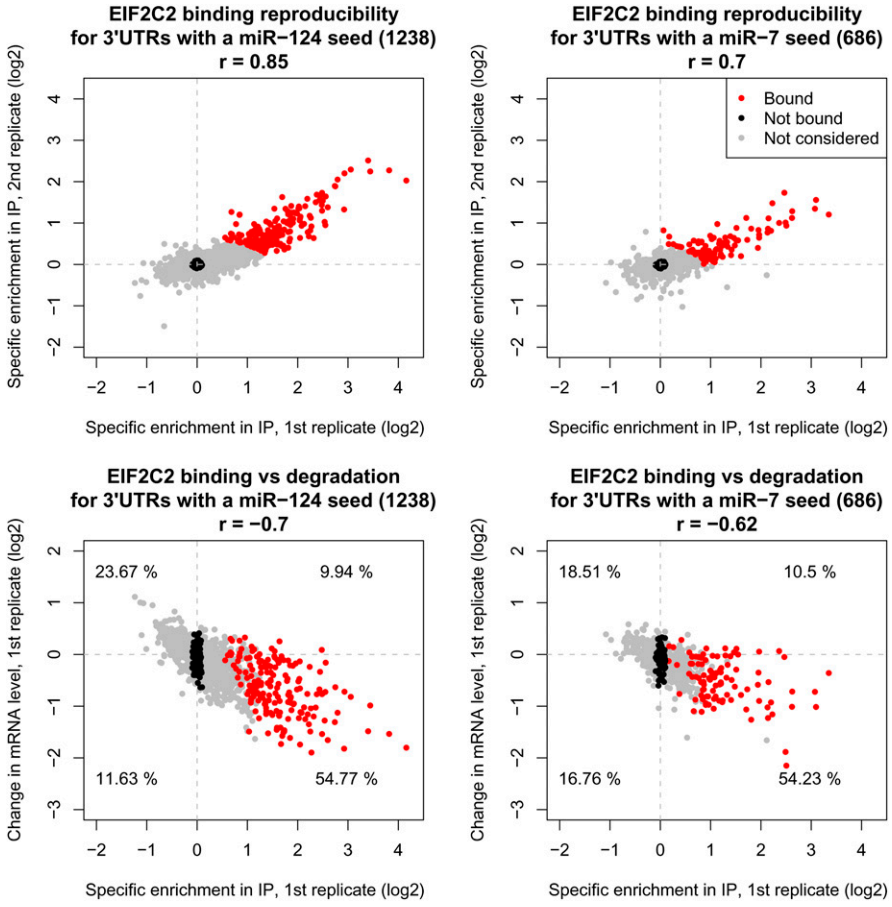

Figure 3. (*Upper* panels) Correlation between the level of EIF2C2 binding in two replicate experiments of transcripts carrying a single seed match for miR-124 (*left*) and miR-7 (*right*) in their 3′ UTRs. The level of EIF2C2 binding was computed as described in the Methods section. The number of transcripts and Pearson correlation coefficients are shown on the respective panels. Transcripts that were considered *positives* for EIF2C2 binding are marked with red, those that were considered *negatives* with black, and transcripts that were not used for feature analysis are shown in gray. (*Lower* panels) Correlation between EIF2C2 binding and mRNA degradation in one of the experiments (miR-124 overexpression in the *left* panel, miR-7 overexpression in the *right* panel). The levels of EIF2C2 binding and mRNA degradation were computed as described in the Methods section. The numbers in the four quadrants indicate the proportion of all transcripts with a single seed-complementary 3′ UTR site that fall in each individual quadrant.

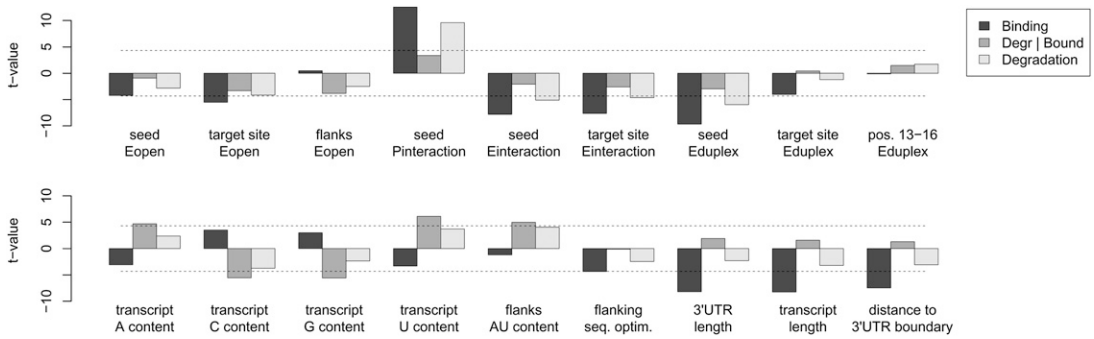

Figure 4. Contribution of secondary structure (*upper* panel), sequence, and transcript length-related (*lower* panel) features to the efficiency of EIF2C2 binding and mRNA degradation. The *y*-axis shows the value of the t-statistic obtained in comparing bound with unbound transcripts (dark gray bars), bound and degraded with bound, but not degraded transcripts (medium gray bars), and degraded with not degraded transcripts (light gray bars). The dashed lines indicate the values beyond which the difference in the mean values obtained for the positive and negative sets is considered significant with a bilateral type I error of 5% after applying the Bonferroni multiple testing correction. The individual features that we tested are indicated and further described in the text.

in nucleotide composition of the regions flanking miRNA target sites are largely a reflection of global biases. Previous studies pointed to the effect of A/U content on the efficacy of miRNA target sites (Jing et al. 2005; Robins and Press 2005; Grimson et al. 2007), although at what step in the miRNA effector cascade this feature plays a role is, so far, unknown. Here we found that this feature comes into play in the degradation of EIF2C2-bound targets. We furthermore found that the U content is more predictive of functionality than the A nucleotide. Interestingly, two known examples of modulation of miRNA activity—the release of miRNA-dependent inhibition of the *SLC7A1 (CAT-1)* mRNA by the ELAVL1 (HuR) protein under stress (Bhattacharyya et al. 2006) and the inhibition of miRNA action in primordial germ cells of zebrafish by DND1 protein (Kedde et al. 2007)—involve interactions of U-rich elements, and a study of mRNA decay (Yang et al. 2003) also identified a number of AU-rich elements that positively correlated with degradation rate.

Consistent with the nucleotide bias, the energy required to open the secondary structure *in the vicinity* of miRNA target sites is lower in the case of functional sites (Fig. 1, "flanks Eopen"). Not all transcriptomics data sets however exhibit this property, which is probably why Grimson et al. (2007) reported that secondary structure prediction was uninformative once the A/U content of the region was taken into account. Interestingly, some of the experiments from Grimson et al. (2007) (Supplemental Fig. 2, miR-133a, miR-142-3p) did not show strong support for structural features, while other experiments in the same series did (Supplemental Fig. 2, miR-122, miR-9).

Implications for target prediction

An immediate question is what features and training sets one should use in order to develop more accurate target prediction methods. To address this question, we constructed three groups of data sets:

The transcriptomics data sets shown in Figure 1, with the exception of the let-7 and miR-30a transfections, that we left out because of the negative feedback on the RNAi pathway.

The proteomics data sets from Selbach et al. (2008) and Baek et al. (2008), again excluding the let-7 and miR-30a experiments.

The comparative genomics data sets from Gaidatzis et al. (2007) that are shown in Figure 1.

Based on a principal component analysis, we selected a set of 14 nonredundant features (Fig. 5, legend) and we trained generalized linear models on the transcriptomics, proteomics, and the combination of the two data sets. In the latter case, we weighted the contribution of the points in the two data sets such that the combined transcriptomics measurements have equal weight in the model as the combined proteomics measurements. Because the extent of evolutionary selection measured by the ElMMo algorithm is a feature in these models, we did not train a model on the comparative genomics data only. We then assessed the predictive power of all three models on all three data sets through receiver operating characteristic (ROC) curves. We added the sensitivities and specificities of some of the current and most distinct target prediction methods for comparison. In cases where models were trained on the same data set, the ROC curve shows the cross-validation specificities and sensitivities. Finally, to get an impression of the upper bound in prediction accuracy that can be expected from a model trained on an experimental data set, we simulated duplicated experiments of varying reproducibility through sampling bivariate Gaussians with correlation coefficients of 0.4 or 0.8. This covers the range of reproducibilities found in the studies whose results we used here, such as the miR-124 transfection of Karginov et al. (2007) (Supplemental Fig. 12).

Unsurprisingly, each of the models that we trained performed very well on the data set on which it was trained. When it comes to predicting transcriptomics data, the model trained on these data performs as well as a replicate experiment with a relatively low (0.4) correlation coefficient would perform (Fig. 5, left panel). In other words, given the noise in some experimental data sets, it is not possible to train a better model from these data, although the situation may change as more reproducible data sets become available. This is illustrated by the comparative genomics ROC curves (Fig. 5, right panel), where it is possible to achieve areas of the curve (AUC) of 0.91, while no model is able to achieve an AUC >0.7 on the transcriptomics or proteomics data sets. On the proteomics data set even the model trained on proteomics only achieves an AUC of 0.6, suggesting that either entirely novel features have to be taken into account in order to explain the protein-level changes that are induced by miRNAs, or that these data sets are too preliminary for studying the determinants of miRNA targeting specificity. Overall, the model trained on transcriptomics data generalizes very well to comparative genomics data, though additional

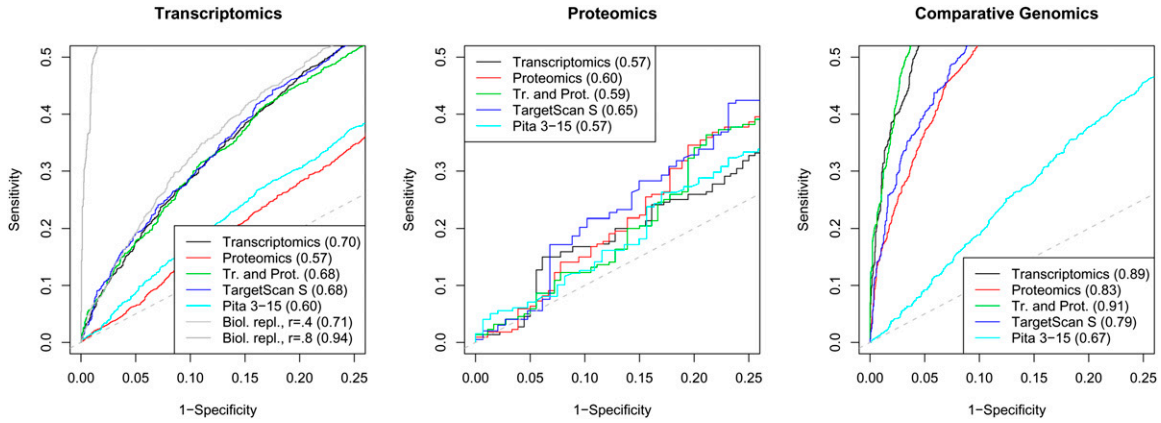

Figure 5. Receiver operating characteristic (ROC) curves of different miRNA target prediction algorithms on transcriptomics, proteomics, and comparative genomics data sets. The numbers that appear in parentheses in the legends indicate the areas *under* the curves (AUC). The model fitted on transcriptomics, proteomics, and combining the transcriptomics and comparative genomics data sets (Tr. and Prot.) include the following features: seed Eopen, target site Eopen, flanks G and U content, 3′ UTR length, ElMMo, seed Pinteraction, seed Eduplex, target site Eduplex, flanking sequence optimization, pos. 13–16 Eduplex, 3′ region Eduplex, distance to 3′ UTR boundary, and relative distance to 3′ UTR boundary. Pita 3–15 is the score according to the algorithm published by Kertesz et al. (2007), and TargetScan is the TargetScan S score from Grimson et al. (2007).

training on the proteomics data still improves slightly the prediction accuracy. Of the previously published models, TargetScanS has good performance on all data sets, which is perhaps due to the fact that it uses features that were inferred from both comparative genomics as well as miRNA transfection and transcriptomics analysis. Interestingly, its performance on the proteomics data set is even better than the performance of the linear model that we trained on the proteomics data itself. On the other hand, the performance of the Pita algorithm (Kertesz et al. 2007) suggests that attempting to predict miRNA targets purely from secondary structure considerations is currently not an optimal strategy.

Discussion

The studies of the determinants of miRNA targeting specificity that have been published so far can be divided into two main classes: those that emphasize sequence features (Robins and Press 2005; Grimson et al. 2007), and those that emphasize mostly structural aspects (Robins et al. 2005; Ameres et al. 2007; Kertesz et al. 2007; Long et al. 2007; Hammell et al. 2008). Because different studies used different systems, looked at different readouts, and had different degrees of precision in the experimental measurements, it has been difficult to reconcile their conclusions concerning the relative importance of these feature in the prediction of miRNA target sites. Here we addressed this problem by applying a uniform battery of tests in order to determine the relative power of individual features in distinguishing functional from nonfunctional target sites. The general conclusion is that a model that combines both sequence as well as structural aspects performs best in miRNA target prediction. The features have nonetheless to be carefully chosen, because the physico-chemistry of miRNA–target interactions is not well characterized at the moment. Thus, although the energy of interaction between a miRNA and its target is generally not a very good predictor, especially when one does not specifically enforce the hybridization of the miRNA seed, structural descriptors improve the predictive power of models that are only based on sequence features. Of the sequence features, we found that the U and A/U content of the 3′ UTRs are the strongest positive and the C and G content of the 3′ UTRs are the strongest negative predictors of miRNA target site functionality (Fig. 2). The question arises of why nucleotide biases computed over regions of the length scale of 3′ UTR lengths are predictive of the functionality of individual sites. One possible answer is that the entire 3′ UTR contributes to the accessibility of individual miRNA binding regions. Consistent with this hypothesis we found that miRNA target site accessibility is one of the strongest structural predictors of target site functionality. On the other hand, we found that target site accessibility is only important for EIF2C2 binding, for which a high A/U content is not predictive. Another possible answer is that various selection pressures act to optimize the nucleotide composition over relatively long regions of the 3′ UTR. This is consistent with the idea that transcripts of certain functional categories, such as transcription factors, are heavily regulated (Robins et al. 2005; Stark et al. 2005), and as a result, their 3′ UTRs are docking platforms for a multitude of regulatory factors, all of which prefer structurally accessible regions. An interesting implication of the length scale of nucleotide compositional biases is that functional target sites will more likely emerge in 3′ UTRs that already have such sites, accompanied by a specific nucleotide bias that extends over long regions. A final possibility is that efficiency of mRNA degradation by exonucleases depends on how extensive the secondary structure of the transcript is. In this scenario, the A/U content of the transcript and its 3′ UTR is not an indicator of the functionality of a miRNA site per se, but sites that are located in A/U-rich transcripts are associated with more efficient target mRNA degradation.

The original paradigm regarding the mechanism of action of miRNAs was that miRNAs cause translational repression of bound mRNAs (Wightman et al. 1993; Reinhart et al. 2000). Further studies have then shown that miRNAs also trigger the degradation of the targeted mRNAs (Bagga et al. 2005; Krützfeldt et al. 2005; Lim et al. 2005), leading to the view that miRNAs primarily cause translation repression, with mRNA degradation occurring as a by-product (Eulalio et al. 2007). Our results here show that the target sites that are under evolutionary selection share most features with the target sites that induce mRNA degradation responses. Thus, we suggest that the translational inhibition only paradigm is the exception rather than the rule, at least for mammalian miRNAs. This conjecture is also supported by the results of EIF2C2-IP and miRNA overexpression/proteomics experiments. The degree of EIF2C2

binding correlates very well with the extent of mRNA degradation (Fig. 3) and there are relatively few targets that appear to be bound by EIF2C2, but not undergo mRNA degradation. Additionally, the proteomics data sets of Selbach et al. (2008) also indicate that there are relatively few targets that appear to be translationally inhibited, yet the corresponding mRNA levels are unchanged (Supplemental Fig. 10). One important exception may be those mRNAs whose translation needs to be inhibited only transiently. Bhattacharyya et al. (2006) described, for instance, the example of the cationic transporter (CAT-1) message, whose inhibition by miR-122 in the liver is reversible under stress. Similar situations arise in neurons, in which the translation of some messages needs to be specifically triggered in response to signals at the neuronal synapse, but not otherwise. For such cases, the measurement of protein levels may be essential in target identification, and it will be extremely interesting to analyze in more depth the targets obtained from proteomics and from transcriptomics measurements performed after transfection of the neuron-specific miRNA, miR-124. Nonetheless, our results indicate that the more common transcriptomics measurements are still very useful for the identification of miRNA targets.

Finally, we found that a model that was trained on transcriptomics data performs better in predicting target sites that are under evolutionary selection than those that are inferred from transcriptomics experiments and that miRNAs appear to feed back on various steps of the RNAi pathway. These findings suggest that a more accurate identification of miRNA target sites may require a deeper quantitative understanding of the miRNA-induced response rather than additional determinants of miRNA targeting specificity.

Methods

microRNA transfection

FLAG/HA-EIF2C2 cells were transfected with miR-7/miR-7* duplex (5′-UGGAAGACUAGUGAUUUUGUUGU/5′-CAACAAAUCACAG UCUGCCAUA) and miR-124/miR-124* duplex (5′-UAAGGCACG CGGUGAAUGCCA/5′-CGUGUUCACAGCGGACCUUGA) or mock and Lipofectamine RNAiMAX as described by the manufacturer (see also Supplemental Fig. 13). Briefly, a 15-cm tissue culture plate was transfected with 900 pmol miRNA duplex and 22 μL Lipofectamine RNAiMAX.

RNA isolation from cell lysate and FLAG-protein immunoprecipitates from FLAG/HA-EIF2C2 expressing cells were lysed in three cell pellet volumes of 50 mM HEPES-KOH at pH 7.4, 150 mM KCl, 2 mM EDTA, 0.5 mM DTT, 1 mM NaF, and 0.5% NP-40. RNA from the lysate was isolated by adding three volumes of RNA extraction solution (4 M guanidinium isothiocyanate, 25 mM Na-citrate, 0.5% N-Lauroylsarcosinate, 50 mM beta-mercaptoethanol, and 50% acidic phenol) and 0.2 volumes of chloroform. RNA was ethanol-precipitated from the aqueous phase. FLAG/HA-tagged EIF2C2 was immunoprecipitated with anti-FLAG M2 agarose beads (Sigma). Beads were washed three times with 50 mM HEPES-KOH at pH 7.4, 300 mM KCl, 2 mM EDTA, 0.5 mM DTT, 1 mM NaF, and 0.05% NP-40. RNA isolation from immunoprecipitated RNPs was performed as described previously (Meister et al. 2004). RNA for microarray analysis was further purified using RNeasy minispin columns (QIAGEN). Quality of the RNA was assessed with the Agilent Bioanalyser.

Dual Luciferase assay of EIF2C2-bound mRNAs

HEK293 cells were cotransfected in 96-well format (40.000 cells/well) with 100 ng of the respective psiCHECK vector and 10 pmol of miRNA duplex or 10 pmol of GFP siRNA duplex (5′-GGCAAGC TGACCCTGAAGTTT/5′-ACTTCAGGGTCAGCTTGCCTT) as control with Lipofectamine 2000 (Invitrogen). Cells were lysed in 1xPassive Lysis Buffer (Promega) 15 h after transfection and analyzed using the Dual-Luciferase Reporter System (Promega) as described by the manufacturer on a BIO-TEK Clarity 96-well plate reader with double injectors.

Dual Luciferase assay of TNRC6A with miR-30a

HeLa cells were transfected in 24-well plates with 5 ng of respective psiCHECK vector using Lipofectamine 2000 (Invitrogen) or cotransfected with 20 nM miR-30a antagomiR (Ambion). Cells were lysed 24 h after transfection and luciferase activities were measured using the Dual-Luciferase Reporter System (Promega) as recommended in the manufacturer instructions.

Microarray experiments

Two micrograms of purified total RNA from HEK293 cell lysate or from immunoprecipitated RNPs were used in the One-Cycle Eukaryotic Target Labeling Assay (Affymetrix) according to the manufacturer's protocol. Biotinylated cRNA targets were then cleaned up, fragmented, and hybridized to Human Genome U133 Plus 2.0 Array (Affymetrix).

Computational analysis of one-channel Affymetrix microarrays from Selbach et al. (2008) and Krützfeldt et al. (2005)

The CEL files of Selbach et al. (2008) were downloaded from http://psilac.mdc-berlin.de/download/ and the antagomiR-122 data of Krützfeldt et al. (2005) was retrieved from the GEO database of NCBI (accession no. GSE3425).

We imported the CEL files into the R software (http://www.R-project.org) using the BioConductor affy package (Gentleman et al. 2004). The probe intensities were corrected for optical noise, adjusted for nonspecific binding, and quantile normalized with the gcRMA algorithm (Wu et al. 2004).

Per gene log2 fold changes were obtained through the following procedure. We first fitted a lowess model of the probe log2 fold change using the probe AU content. We used this model to correct for the technical bias of AU content on probe-level log2 fold change reported by Elkon and Agami (2008). Subsequently, probe set-level log2 fold changes were defined as the median probe-level log2 fold change. Probe sets with more than two probes mapping ambiguously (more than one match) to the genome were discarded, as were probe sets that mapped to multiple genes. We then collected all remaining probe sets matching a given gene, and averaged their log2 fold changes to obtain an expression change per gene. For sequence analyses, we selected for each gene the RefSeq transcript with median 3′ UTR length corresponding to that gene.

Finally, we considered all genes for which at least one probe set was called present in the transfection experiments as expressed, and went on analyzing only these genes while ignoring all other genes.

Computational analysis of two-channel Agilent microarrays from Karginov et al. (2007) and Baek et al. (2008)

The Baek data set was downloaded from the GEO database of NCBI (accession no. GSE11968). For the Karginov data set we started

with the text file output of the Agilent scanner, which was kindly provided to us by Ted Karginov.

We extracted the rProcessedSignal, gProcessedSignal, LogRatio, rIsWellAboveBG, and gIsWellAboveBG fields for each probe, keeping only probes for which both gIsWellAboveBG and rIsWellAboveBG flags were true in all experiments. We then quantile normalized the green and red channel intensities, which we obtained from the rProcessedSignal and gProcessedSignal fields of all experiments together. We computed probe-level log2 fold changes from the quantile-normalized rProcessedSignals and gProcessedSignals.

After discarding probes mapping to multiple genes, we collected all probes matching a given gene, and we estimated the log2 fold change per gene as the average log2 fold change of the probe sets associated with it. Finally, for each gene we selected for further sequence analysis the RefSeq transcript with median 3′ UTR length corresponding to that gene.

Computational analysis of two-channel Agilent microarrays from Linsley et al. (2007) and Grimson et al. (2007)

We downloaded the processed differential expression data from GEO (accession nos. GSE6838 and GSE8501) together with the probe to transcript mapping provided by the authors as a SOFT formatted file. For subsequent analysis, we kept only probes associated to RefSeq transcripts according to the annotation. We used all the experiments in the Grimson et al. (2007) series. From the microarray data provided by Linsley et al. (2007), we kept only those experiments that had quasi-replicates (transfections in both HCT116 and DLD-1 cells). These involved let-7c, miR-103, miR-106b, miR-141, miR-15a, miR-16, miR-17, miR-192, miR-200a, miR-20a, and miR-215 transfections and microarray measurements at 24h (GEO accession nos. GSM156546, GSM156550, GSM156545, GSM156549, GSM156543, GSM156576, GSM156532, GSM156541, GSM156534, GSM156542, GSM156580, GSM156544, GSM156547, GSM156551, GSM156548, GSM156552, GSM156553, GSM156555, GSM156554, GSM156556, GSM156557, and GSM15655).

Computational analysis of SILAC assay from Baek et al. (2008)

We downloaded the data provided by the authors in the Supplemental Material and used it without any specific post-processing.

Computational analysis of pSILAC assay from Selbach et al. (2008)

We downloaded the "all peptide evidence" flat file from http://psilac.mdc-berlin.de/download/.

We mapped all peptides in the pSILAC data set against the RefSeq Protein database from August 14th 2008 using wu-blastp 2.0 and a seed word length of 5, discarding alignments with gaps or with more than one mismatch. We further discarded peptides that mapped to more than one protein.

Per protein log2 fold changes were computed for all proteins credited with 3–15 peptides log2 fold changes across replicates and gel slices.

EIF2C2 binding affinities in the Karginov data set

Transcript degradation was quantified as the logarithm of the ratio of transcript expression in the lysates of miRNA-transfected and mock-transfected cells. The miRNA-specific EIF2C2 binding was quantified as the ratio of two ratios: EIF2C2-IP of miRNA-transfected and mock-transfected cells and lysates of miRNA-transfected and mock-transfected cells (Supplemental Fig. 14).

Extraction of positives and negatives from replicated transfection experiments

Among the transcriptomics data sets we reanalyzed, the experiments performed by Grimson et al. (2007), Selbach et al. (2008), Baek et al. (2008) and Krützfeldt et al. (2005) did not feature biological replicates. For these data sets, we considered the top 250 down-regulated (or up-regulated for Krützfeldt et al. [2005]) transcripts that carried precisely one seven-mer or eight-mer seed match. After discarding all seed matches located in the CDS, we ended up with a set of *positive* seed matches. The negatives were obtained through selecting the 250 least-changing transcripts with seed matches, that is the 250 transcripts whose log2 expression fold changes were closest to 0 when comparing the miRNA-transfected samples to the mock-transfected samples. After discarding all seed matches located in the CDS, we ended up with a set of *negative* seed matches.

The experiments performed by Linsley et al. (2007) and Karginov et al. (2007), on the other hand, featured biological replicates. For these data sets, we applied a method that we designed for selecting transcripts that, with high probability, are affected in expression by the miRNA across all experiments in which the expression of the given miRNA was perturbed (see Supplemental Material). Briefly, we first need to calculate, for each pairwise microarray comparison (further referred to as contrast) k, the probability $P_k(f|-)$ that a transcript that is not a target, will have a log fold change of f. To estimate the distributions $P_k(f|-)$ we assumed that they are Gaussian with means μ_k and standard deviation σ_k to be estimated from the data for each contrast k. In addition we assumed that transcripts that do not carry at least a heptameric seed-complementary site are unlikely to be real targets, and thus estimated μ_k and σ_k from the observed expression changes of transcripts without such seed matches. We similarly need to calculate, for each contrast k, a distribution $P_k(f|+)$ that a transcript, which is a true target of the miRNA, will have fold-change f. As little is currently known of the distribution of the severity of the effect that miRNAs have on the expression of their targets we assumed as little as possible about the distribution $P_k(f|+)$, namely that a true target must change expression in the right direction, i.e., $f<0$ for a miRNA overexpression experiment, and $f>0$ for a miRNA knockdown experiment, and that expression changes are limited to a finite range over which the expression change has a *uniform* distribution. Finally, based on these distributions, we estimate the posterior probability that a transcript with fold change f is a functional target in a given experiment. Details are given in the Supplemental Material. The same procedure was used to construct the sets of positives and negatives from our miR-124 and miR-7 transfection experiments. The process is illustrated in Supplemental Figure 15 and the lists of transcripts with a posterior probability of ≥ 0.5 of being functional in both contrasts of our two miRNA transfection experiments are shown in Supplemental Tables 1 and 2. For the negatives we selected those transcripts with minimal sum of squared log2 fold changes in the two experiments. Finally, for the feature analysis, we then proceeded as with experiments where no replicates were performed: we selected 250 positives and 250 negatives according to the criteria defined above and we discarded those cases in which the seed match was in the CDS.

Extraction of positives and negatives from EIMMo predictions

From our predictions of miRNA target sites inferred to be under evolutionary selection (Gaidatzis et al. 2007) and for each of the experimentally tested and conserved miRNAs (miR-30a, let-7c, miR-155, miR-1, miR-103, miR-15a, miR-16, miR-106b, miR-20a, miR-141, miR-200a, miR-181a, miR-124, and miR-17), we selected

the top 250 target sites in the order of their posterior probability of being under selection. We also selected an equal number of sites least likely to be under selection.

Feature definition and computation

To minimize the ambiguity of attributing a specific response to a miRNA binding site, we only analyzed transcripts that had precisely one miRNA seed match (complementarity to positions 1–7, 2–8, or 1–8 of the miRNA) and the site was at least 100 nt away from either of the boundaries of the 3′ UTR. A sketch of the transcript regions used for the various computations below is shown in Supplemental Figure 16. For each individual putative target site we then computed the following quantities.

Seed accessibility (seed Eopen) was defined in terms of the energy necessary to open the secondary structure of the target in the region binding positions 1–8 of the miRNA. This was computed using the program RNAup of the Vienna package (Hofacker 2003) with the following parameters: $u = 8$ (length of the window required to be single-stranded), $w = 50$ (maximal length of the interacting region). The rest of the parameters were left with their default values. Other choices of the w parameter did not qualitatively affect our results (not shown). The negative value of this energy can be viewed as a measure of accessibility.

Site accessibility (site Eopen) was similarly defined in terms of the energy required to open the secondary structure of the target in a region of 20 nt, anchored at the 3′ end by the seed-complementary region (opposite positions 1–8 of the miRNA). The computation was performed as described above, except that we used a window size u of 20 instead of 8.

Accessibility of the flanks (flanks Eopen) was defined as the average accessibility (defined above) of a window of length 20 contained in the regions of 50 nt upstream or 50 nt downstream of the miRNA target site.

Seed hybridization energy (seed Eduplex) is the energy ΔG_h of the hybrid formed between the seed (positions 1–8 of the miRNA) and the seed-complementary site, as given by the RNAduplex program of the Vienna package (Hofacker 2003).

MiRNA hybridization energy (target site Eduplex) is the energy of the hybrid formed between the miRNA (positions 1–20) and the miRNA-complementary site, as given by the RNAduplex program.

Contribution of different 3′ regions of the miRNA to the hybridization (pos. 9–12 Eduplex, pos. 13–16 Eduplex, pos. 13–20 Eduplex) is the difference $\Delta G_u - \Delta G_c$ between the minimum binding free energy ΔG_h of the full mRNA-miRNA duplex and the binding free energy ΔG_c of the same duplex under the constraint that nucleotides 9–12, 13–16, or 13–20 of miRNA are unpaired, respectively. The duplex structure with minimum binding free energy was computed by the RNAduplex program of the Vienna package. Starting from this structure, we enforced the constraints at positions 9–12, 13–16, and 13–20 and computed the corresponding binding free energy ΔG_c using RNAeval from the Vienna package (Hofacker 2003).

Seed interaction energy (seed Einteraction) was defined as $\Delta G = \Delta G_o + \Delta G_h$, where ΔG_o is the energy required to open the secondary structure of the target in the seed-complementary region and ΔG_h is the energy of the hybrid formed between the seed and the seed-complementary site. ΔG_o is obtained as described in the paragraph "Seed accessibility" above, and ΔG_h is computed using the RNAduplex program (Hofacker 2003) with default parameters. Note that we neglected the energy, possibly required, to open the structure of the seed region of the miRNA. The probability of interaction with the seed region of the miRNA (seed Pinteraction) is the corresponding probability, as computed by RNAup.

MiRNA interaction energy (target site Einteraction) was similarly defined as $\Delta G = \Delta G_o + \Delta G_h$, where ΔG_o is the energy required to open the secondary structure of the target in the miRNA-binding region of 20 nt anchored at the seed (as described above) and ΔG_h is the energy of the hybrid formed between the miRNA and the miRNA-complementary site. ΔG_o is obtained as described above, and ΔG_h is computed using the RNAduplex program (Hofacker 2003) with default parameters.

Flanks A, C, G, and U contents were defined as the proportions of A, C, G, and U nucleotides within 50 nt upstream and 50 nt downstream of the miRNA binding site of 20 nt, anchored downstream by the seed-matching region.

3′ UTR A, C, G, and U contents were defined as the proportions of A, C, G, and U nucleotides within the 3′ UTR harboring the miRNA binding site.

Transcript A, C, G, U, and AU contents were defined as the proportions of A, C, G, U, and A + U nucleotides in the transcript harboring the miRNA binding site.

Transcript and 3′UTR length were obtained from the RefSeq sequence and annotation.

Relative position was computed by dividing the position in the 3′ UTR marking the beginning of the seed complementary region by the 3′ UTR length.

Relative distance to 3′UTR boundary was computed similarly, dividing the minimal distance from the beginning of the seed complementary region to the STOP codon or the poly-A tail by the length of the 3′ UTR.

Flanking sequence optimization was designed to measure the extent to which the nucleotide composition of the regions flanking a miRNA binding site explains the accessibility of the miRNA binding site. For each target site we generated 100 variants in which we randomized, independently of each other, the sequence of the 50 nt upstream and of the 50 nt downstream of the miRNA target site, while keeping the mononucleotide frequencies in these regions constant. For the randomized variants we recomputed the accessibility of the miRNA binding site as described above. We then calculated the z-statistic of the real sequence relative to the randomized variants. This computation gave us one set of z-statistics for the positives and one for the negatives. We finally used the t-test to compare the means of the two distributions of z-statistics.

ElMMo is the posterior probability that a seed complementary region is under evolutionary selective pressure described in Gaidatzis et al. (2007).

Testing different linear models for predicting various types of miRNA target sites

We divided all the data sets that we studied here into three groups, as described in the Implications for Target Prediction section. We then performed a principal component analysis to determine a set of 14 nonredundant features (Fig. 5, legend). We then used these features to train three independent generalized linear models (GLM) with logit link function (McCullagh and Nelder 1989) on the transcriptomics data sets, on the proteomics data sets, and a mixture of the transcriptomics and proteomics data sets. In the latter case, we weighted each putative miRNA target site proportionally to the inverse of the data set size, to have the resulting model minimize the prediction error equally on both data sets.

To avoid overestimating the performance of the three GLMs when testing them on the data sets on which they were trained, we performed 10-fold cross-validation. In other words, we split our data set into 10 parts, trained the model using the first nine parts of the data set, and evaluated its sensitivity and specificity on the last. We reiterated this procedure 10 times and used the numbers that

came out of it to plot the cross-validated receiver operating characteristic (ROC) curve. The ROC curves for GLMs trained on different data sets and for other miRNA target prediction algorithms were computed using the standard procedure (Spackman 1989).

To simulate ROC curves from biological replicates of varying reproducibilities, we sampled 25,000 points from bivariate Gaussians with correlation coefficients r of 0.4 and 0.8, which covers the range of reproducibilities of log2 fold changes that we observed in the experimental data sets that we analyzed here. We then considered the 10% (2500) smallest values from the first simulated replicate as fold changes in a transfection experiment for "true target sites" and attempted to use the second simulated replicate to predict the true target sites. The two ROC curves show to what extent knowing one simulated data set enables one to predict the other depending on whether the replicates are in moderate ($r = 0.4$) or good agreement with each other ($r = 0.8$).

Evaluating the competition between the endogenous and the transfected miRNA

Khan et al. (2009) recently reported that transfected miRNAs compete with the endogenous miRNAs for RISC loading. To evaluate this effect in the context of our study, we applied the analysis methods of Khan et al. (2009) to all 44 microarrays performed in the HCT116 Dicer −/−, 8, 10, 14, and 24 h after miRNA or siRNA transfection (GEO accession nos, GSM156513, GSM156514, GSM156515, GSM156516, GSM156517, GSM156518, GSM156519, GSM156520, GSM156567, GSM156568, GSM156569, GSM156570, GSM156571, GSM156572, GSM156573, GSM156574, GSM156525, GSM156526, GSM156527, GSM156536, GSM156521, GSM156522, GSM156523, GSM156524, GSM156531, GSM156532, GSM156533, GSM156534, GSM156545, GSM156546, GSM156547, GSM156548, GSM156553, GSM156554, GSM156557, GSM156559, GSM156565, GSM156566, GSM156575, GSM156576, GSM156577, GSM156578, GSM156579, GSM156580, and GSM156581) and to microarrays that monitored the mRNA expression changes at 8 and 32 h after the transfection of five miRNAs published by Selbach et al. (2008).

To be able to compare our results with those of Khan et al. (2009), we slightly modified the microarray data processing described in the Computational Analysis of One-Channel Affymetrix Microarrays from Selbach et al. (2008) and Krützfeldt et al. (2005) section and the Computational Analysis of Two-Channel Agilent Microarrays from Linsley et al. (2007) and Grimson et al. (2007) section: At the step where we choose a representative RefSeq mRNA for each gene monitored on the microarray we chose the RefSeq mRNA with the longest 3′ UTR, instead of the RefSeq with the median-length 3′ UTR.

We determined the set "X" of mRNAs whose 3′ UTRs carried a match to positions 2–8 of the transfected miRNA. We then determined the set "D" of mRNAs carrying a 2–8 seed match to one of the top 10 miRNA families most expressed in the cell line (HCT116 Dicer −/− or HeLa) used in the experiment. We used the miRNA family expression profiles reported in Supplemental Figure 2 of Khan et al. (2009). We determined the set "B" of mRNAs that carried seed matches to neither the transfected miRNA nor the top 10 endogenous miRNA families. Finally, we applied a linear transformation to the log2 fold change such that the log fold changes of the mRNA belonging to the B set had a mean of 0 and a variance of 1.

We then computed the average log fold changes of the X, $X \cap D$, $X \setminus D$, $D \setminus X$ and B mRNA sets. Doing so for each set of mRNAs, and for each time point, gave us one measurement of mRNA log fold change per experiment (i.e., per transfected miRNA), which we combined by averaging over all experiments performed at the same time point and computing the 95% confidence interval on the mean log fold changes.

Acknowledgments

We thank Thomas Tuschl for support in performing the experiments, and for suggestions and discussions. We are grateful to Wenxiang Zhang and Connie Zhao (Rockefeller University Genomics Resource Center) for mRNA array analyses, to Erik van Nimwegen (Biozentrum, University of Basel) for suggesting the model to identify functional sites from multiple experiments, and to Ted Karginov for providing us the raw Agilent data of the miR-124 EIF2C-IP. J.H. was supported by the Swiss National Fund, grant no. 3100A0-114001, to Mihaela Zavolan. M.L. was partially supported by an Irma T. Hirschl Postdoctoral Fellowship. In addition, M.L. was supported by NIH grant no. GM068476 to Thomas Tuschl. We also thank Lukas Burger, Erik van Nimwegen, and Walter Keller (Biozentrum, University of Basel) for critical comments on the manuscript.

References

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

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

Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R, Pasquinelli A. 2005. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122:** 553–563.

Bhattacharyya S, Habermacher R, Martine U, Closs E, Filipowicz W. 2006. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell* **125:** 1111–1124.

Brennecke J, Stark A, Russell R, Cohen S. 2005. Principles of microRNA–target recognition. *PLoS Biol* **3:** e85. doi: 10.1371/journal.pbio.0030085.

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

Elkon R, Agami R. 2008. Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs. *PLoS Comput Biol* **4:** e1000189. doi: 10.1371/journal.pcbi.1000189.

Enright A, John B, Gaul U, Tuschl T, Sander C, Marks D. 2003. MicroRNA targets in *Drosophila*. *Genome Biol* **5:** R1.

Eulalio A, Behm–Ansmant I, Izaurralde E. 2007. P bodies: At the crossroads of post-transcriptional pathways. *Nat Rev Mol Cell Biol* **8:** 9–22.

Forman J, Legesse-Miller A, Coller H. 2008. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci* **105:** 14879–14884.

Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. 2007. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* **8:** 69. doi: 10.1186/1471-2105-8-69.

Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80. doi: 10.1186/gb-2004-5-10-r80.

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

Hammell M, Long D, Zhang L, Lee A, Carmack C, Han M, Ding Y, Ambros V. 2008. mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat Methods* **5:** 813. doi: 10.1038/nmeth.1247.

Hendrickson D, Hogan D, Herschlag D, Ferrell J, Brown P. 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.

Hofacker I. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* **31:** 3429–3431.

Jing Q, Huang S, Guth S, Zarubin T, Motoyama A, Chen J, Di Padova F, Lin S, Gram H, Han J, et al. 2005. Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell* **120:** 623–634.

Karginov F, Conaco C, Xuan Z, Schmidt B, Parker J, Mandel G, Hannon G. 2007. A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci* **104:** 19291–19296.

Kedde M, Strasser M, Boldajipour B, 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.

Khan AA, Betel D, Miller ML, Sander C, Leslie CS, Marks DS. 2009. Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nat Biotechnol* **27:** 549–555.
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 E. 2002. Micro RNAs 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 P, Soll S, 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 L, Lau N, Garrett–Engele P, Grimson A, Schelter J, Castle J, Bartel D, Linsley P, Johnson J. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–773.
Linsley P, Schelter J, Burchard J, Kibukawa M, Martin M, Bartz S, Johnson J, Cummins J, Raymond C, 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 W, Ohler U. 2007. Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics* **8:** 152. doi: 10.1186/1471-2164-8-152.
McCullagh P, Nelder JA. 1989. *Generalized linear models*. Chapman & Hall/CRC, London.
Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15:** 185–197.
Reinhart B, Slack F, Basson M, Pasquinelli A, Bettinger J, Rougvie A, Horvitz H, Ruvkun G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403:** 901–906.
Robins H, Press W. 2005. Human microRNAs target a functionally distinct population of genes with AT-rich 3′ UTRs. *Proc Natl Acad Sci* **102:** 15557–15562.
Robins H, Li Y, Padgett R. 2005. Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci* **102:** 4006–4009.
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.
Spackman KA. 1989. Signal detection theory: Valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pp. 160–163.
Stark A, Brennecke J, Russell R, Cohen S. 2003. Identification of *Drosophila* microRNA targets. *PLoS Biol* **1:** e60. doi: 10.1371/journal.pbio.0000060.
Stark A, Brennecke J, Bushati N, Russell R, Cohen S. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123:** 1133–1146.
Tafer H, Ameres S, Obernosterer G, Gebeshuber C, Schroeder R, Martinez J, Hofacker I. 2008. The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol* **26:** 578–583.
Tokumaru S, Suzuki M, Yamada H, Nagino M, Takahashi T. 2008. let-7 regulates Dicer expression and constitutes a negative feedback loop. *Carcinogenesis* **29:** 2073–2077.
Wightman B, Ha I, Ruvkun G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855–862.
Wu Z, Irizarry R, Gentleman R, Martinez-Murillo F, Spencer F. 2004. A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* **99:** 909–917.
Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell J. 2003. Decay rates of human mRNAs: Correlation with functional characteristics and sequence attributes. *Genome Res* **13:** 1863–1872.
Zhao Y, Ransom J, Li A, Vedantham V, von Drehle M, Muth A, Tsuchihashi T, McManus M, Schwartz R, Srivastava D, et al. 2007. Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell* **127:** 303–317.

Received January 12, 2009; accepted in revised form July 8, 2009.