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,*], Markus Landthaler[2,*,‡], Lukasz Jaskiewicz[1], Dimos Gaidatzis[1,†], Mihaela Zavolan[1,§]

July 8, 2009

[1]Biozentrum, University of Basel and Swiss Institute of Bioinformatics Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland.

[2]Howard Hughes Medical Institute, Laboratory for RNA Biology, The Rockefeller University, New York, 10021 NY, USA.

[†]Current address: Friedrich-Miescher Institute for Biomedical Research, CH-4058 Basel, Switzerland.

[‡]Current address: Max-Delbrück-Centrum für Molekulare Medizin (MDC) Berlin-Buch, Robert-Rössle-Strasse 10, 13125 Berlin-Buch, Germany.

[*]These authors contributed equally to this work.

[§]Correspondence to: mihaela.zavolan@unibas.ch.

1 Supplementary Material

1.1 Plasmids and cell culture

FlPin293 cells stably expressing FLAG/HA *EIF2C2* were described in Landthaler et al. (2008).

The DNA oligonucleotides used for the amplification of 3' UTRs were the following (restriction sites are underlined):

CHMP4A-1: 5'-CCGCTCGAGTAAATCTGGGCTTGTCTTCCTAATGCTACC,

CHMP4A-2: 5'-GAATGCGGCCGCGGGAACAAGGGCATTATAACTGCTATCAAAG;

LOC134145-1: 5'-CCGCTCGAGACTTGACTGGGAGTGTTTTTCTGAAATATTGTAG,

LOC134145-2: 5'-GAATGCGGCCGCAAGTTTAGTTAAAGATGTGACCATCTTACTTCATTAC;

AXIN1-1: 5'-CCGCTCGAGCAAAGTGGAGAAGGTGGACTGATAG,

AXIN1-2: 5'-GAATGCGGCCGCTCATTATTATCCAAGTACCTTTGAAAAGATAATTAATTG;

ANXA5-1: 5'-CCGCTCGAGTGTCACGGGGAAGAGCTCCCTG,

ANXA5-2: 5'-GAATGCGGCCGCTCATTAATCTTTTGAATACAATCATCATAATTTTACAGG;

KANK1-1: 5'-CCGCTCGAGTATGCAAATAGCCCTTTATTTACATGCCAC,

KANK1-2: 5'-GAATGCGGCCGCTTTGAAAATATGGCAAGAGTCTAAGGCACTTC;

PGRMC2-1: 5'-CCGCTCGAGACTTTGTAAACAACCAAAGTCAGGGGCCTTC

PGRMC2-2: 5'-GAATGCGGCCGCGTACATGCTTTATTAAAATGGTACTTGTATTTACAG;

RNF128-1: 5'-CCGCTCGAGTCTGTGTAAATAGAAAACTTGAACCATTAGTAATAAC,

RNF128-2: 5'-GAATGCGGCCGCACATTTTATATTTAAAGAGAATCAATACAAATTGGGAC.

1.2 Extraction of positives and negatives from replicated transfection experiments

For the set of "positives" we wanted to select transcripts that, with high probability, are affected in expression across all experiments in which the expression of a miRNA was perturbed. We therefore developed a probabilistic model that, for each transcript containing one or more miRNA seed matches, uses the expression data from over-expression or knock-down experiments of the corresponding miRNA, to calculate the probabilities that the transcript's expression is affected by the miRNA in each of these experiments.

For the purpose of this model, we define a miRNA seed match as a 7mer or 8mer perfect match to the miRNA seed. We assume that our data consists of K pairs of expression measurements, each corresponding to either a miRNA over-expression or miRNA knock-down experiment, which we will refer to as "contrasts". We will let f_t^k denote the $\log_2$ fold-change of expression of transcript t in contrast k.

Distribution of fold-changes for non-targets

For our model we first need to calculate, for each 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. We in addition 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 seed matches.

Distribution of fold-changes for targets

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 a 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 will assume as little as possible about the distribution $P_k(f|+)$. The only thing that we will assume is that a true target must change expression in the right direction, i.e. $f < 0$ for a miRNA over-expression experiment, and $f > 0$ for a miRNA knock-down experiment, and that expression changes are limited to a finite range. That is, we will assign a *uniform* distribution. For example, in the case of a contrast related to a miRNA over-expression:

$$P_k(f|+) = \begin{cases} \frac{1}{|F_k|} & \text{if } F_k \leq f < 0 \\ 0 & \text{otherwise} \end{cases}$$

where $F_k = \min_t(f_t^k)$ is the largest negative $\log_2$ fold-change observed in contrast k. The distribution is defined in a similar fashion for contrasts related to a miRNA knock-down, except it is uniform over *positive* instead of negative values.

Computing the probability of a functionality pattern given the data

The simplest assumption that one could make is that each transcript t is either a true target in each contrast or not a target in any of the contrasts. However, inspection of the data strongly suggested that a transcript t can show a strong response in some experiments and no responses in others. Therefore we developed a more general model in which a transcript can be a "functional target" in some experiments and a non-target in other experiments. We define a functionality pattern α as $\alpha \in S := \{+,-\}^K$. For instance, $\alpha = (\alpha_1, \alpha_2) = (-,+)$ means that the transcript is not a functional target in the first contrast but it is a functional target of the miRNA in the second contrast.

Let D be the whole set of microarray data $D := \{f_t^k\}_{t=\{1,\ldots,T\},k=\{1,\ldots,K\}}$, with T being the number of transcripts and K the number of contrasts. Let further D_t be the microarray data we have about transcript t, $D_t := \{f_t^k\}_{k=\{1,\ldots,K\}}$.

Consider the case where we have $K = 2$ contrasts. What would like to compute ultimately is the posterior probability that a transcript t, which is harboring a seed match (transcripts without seed matches are assumed non-targets per definition), is a functional target of the miRNA whose expression we perturbed in the 2 experiments given the observed $\log_2$ expression fold changes f_t^1, f_t^2. Using Bayes' theorem, we have

$$P(+,+|f_t^1, f_t^2) = \frac{P(f_t^1, f_t^2|+,+)\rho_{++}}{\sum_{\alpha \in S} P(f_t^1, f_t^2|\alpha)\rho_\alpha}.$$

Here we have introduced the *prior* probabilities ρ_α which give the probabilities that a randomly chosen transcript with a seed match will have functionality pattern α. For example ρ_{++} is the prior probability that a randomly chosen transcript with seed match is functional in both contrasts. As shown below, the

ρ_α are unknown parameters which we set by maximizing the probability of the data D.

Fitting prior probabilities

Under our model, the probability of the observed fold-changes D_t for a given transcript t is given by

$$P(D_t|\rho) = \sum_\alpha \rho_\alpha \prod_{k=1}^{K} P_k(f_t^k|\alpha_k) = \sum_\alpha \rho_\alpha P(D_t|\alpha), \tag{1}$$

where α_k is the k-th component of the functionality pattern α (either $-$ or $+$), and we have defined the probability $P(D_t|\alpha)$ of the data D_t given pattern α in the last equality. The probability of the entire data set is simply the product over all transcripts t:

$$P(D|\rho) = \prod_{t=1}^{T} \left[\sum_\alpha \rho_\alpha P(D_t|\alpha) \right]. \tag{2}$$

We now want to maximize $P(D|\rho)$ with respect to the prior probabilities ρ_α while satisfying the constraint $\sum_\alpha \rho_\alpha = 1$. This can be done using the method of Lagrange multipliers. We let $L(\rho)$ denote the log-likelihood of the parameters ρ, i.e.

$$L(\rho) = \log \left[P(D|\rho) \right]. \tag{3}$$

The optimal ρ_α then satisfy the following equations

$$\frac{\partial L(\rho)}{\partial \rho_\alpha} = c \; \forall \alpha, \tag{4}$$

where c is a constant (the Lagrange multiplier).

We find for the derivative of the log-likelihood

$$\frac{\partial L(\rho)}{\partial \rho_\alpha} = \sum_{t=1}^{T} \frac{P(D_t|\alpha)}{\sum_\beta \rho_\beta P(D_t|\beta)}. \tag{5}$$

From the above equation it is easy to see that

$$\sum_\alpha \rho_\alpha \frac{\partial L(\rho)}{\partial \rho_\alpha} = T. \tag{6}$$

Combining this with equation (4) we find that the Lagrange multiplier is given by

$$c = T \tag{7}$$

and from this it follows that, at the optimum, the ρ_α satisfy:

$$\rho_\alpha = \frac{1}{T} \sum_{t=1}^{T} \frac{P(D_t|\alpha)\rho_\alpha}{\sum_\beta \rho_\beta P(D_t|\beta)}. \tag{8}$$

We can solve these equations using an Expectation-Maximization (EM) procedure. We start with a random distribution ρ and use the above equation as an update equation, i.e. at each iteration with replace ρ with $\tilde{\rho}$ according to the equation

$$\tilde{\rho}_\alpha = \frac{1}{T} \sum_{t=1}^{T} \frac{P(D_t|\alpha)\rho_\alpha}{\sum_\beta \rho_\beta P(D_t|\beta)}, \tag{9}$$

until the distribution no longer changes. It is easy to show that the second derivatives of the log-likelihood are all negative, i.e.

$$\frac{\partial^2 L(\rho)}{\partial \rho_\alpha \partial \rho_\beta} \leq 0 \; \forall \alpha, \beta. \tag{10}$$

Therefore, the log-likelihood $L(\rho)$ is a convex function and the EM procedure will lead to the unique global optimum which we will denote by ρ^*.

Posterior probabilities of functionality

Using the fitted priors ρ^*_α we can now calculate, for each transcript t, the posterior probabilities $P(\alpha|D_t)$ that it has functionality pattern α. Using Bayes' theorem we have

$$P(\alpha|D_t) = \frac{P(D_t|\alpha)\rho^*_\alpha}{\sum_\beta P(D_t|\beta)\rho^*_\beta}. \tag{11}$$

In particular, for the cases where there are two contrasts (like in our data) we can calculate the posterior probabilities $P(++|D_t)$ that the transcript is functional in both contrasts (see Supplementary Figure 15). We sorted all transcripts by this probability $P(++|D_t)$ and selected the positives as the top n transcripts in this list.

Negatives

On the other hand, as negatives we wanted to select transcripts that behave consistently, i.e. not responding, in replicated experiments. We therefore computed the sum of squared log2 fold changes of each transcript in the two experiments and we chose a number of transcripts matching the number of positives starting from the lowest to the highest value.

References

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

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A., Landthaler, M., *et al.*, 2007. A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell*, **129**:1401–1414.

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

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.

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

2 Supplementary Figures

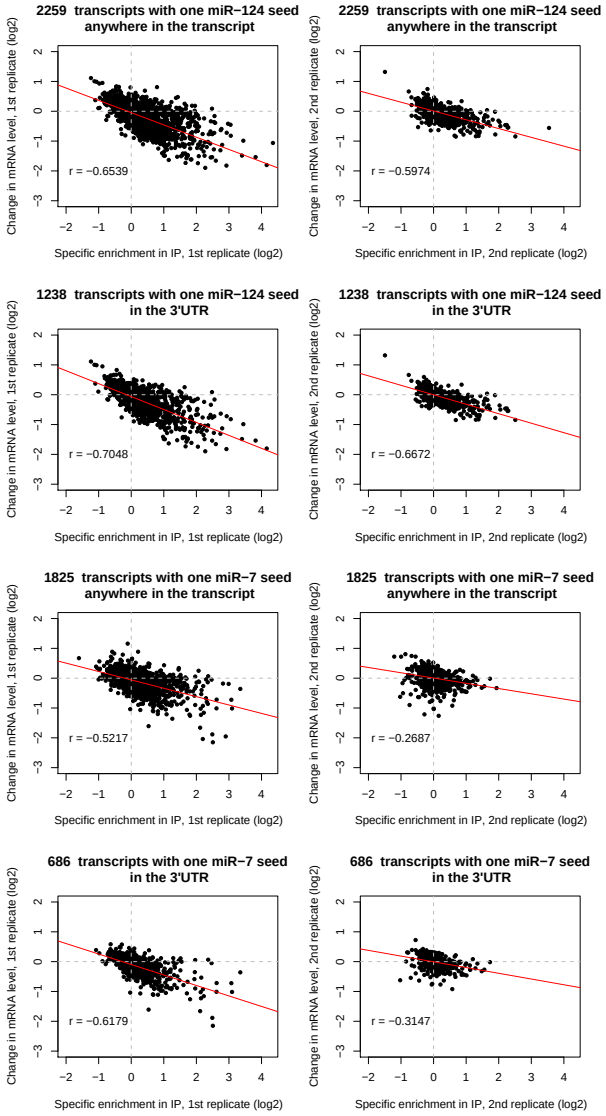


Figure 1: Correlation between the degree of *EIF2C2* binding and the extent of mRNA degradation in transcripts in which the single miRNA seed-complementary site is located in the 3'UTR or anywhere in the transcript for the miR-124 and miR-7 *EIF2C2*-IP. Each row shows a given comparison for the replicate experiments: miR-124 seed match anywhere in the transcript, miR-124 seed match in 3' UTR, miR-7 seed match anywhere in the transcript, miR-7 seed match in 3' UTR. The values of the Pearson correlation coefficients are indicated on the plots and the number of transcripts used for each plot is indicated in the corresponding title.

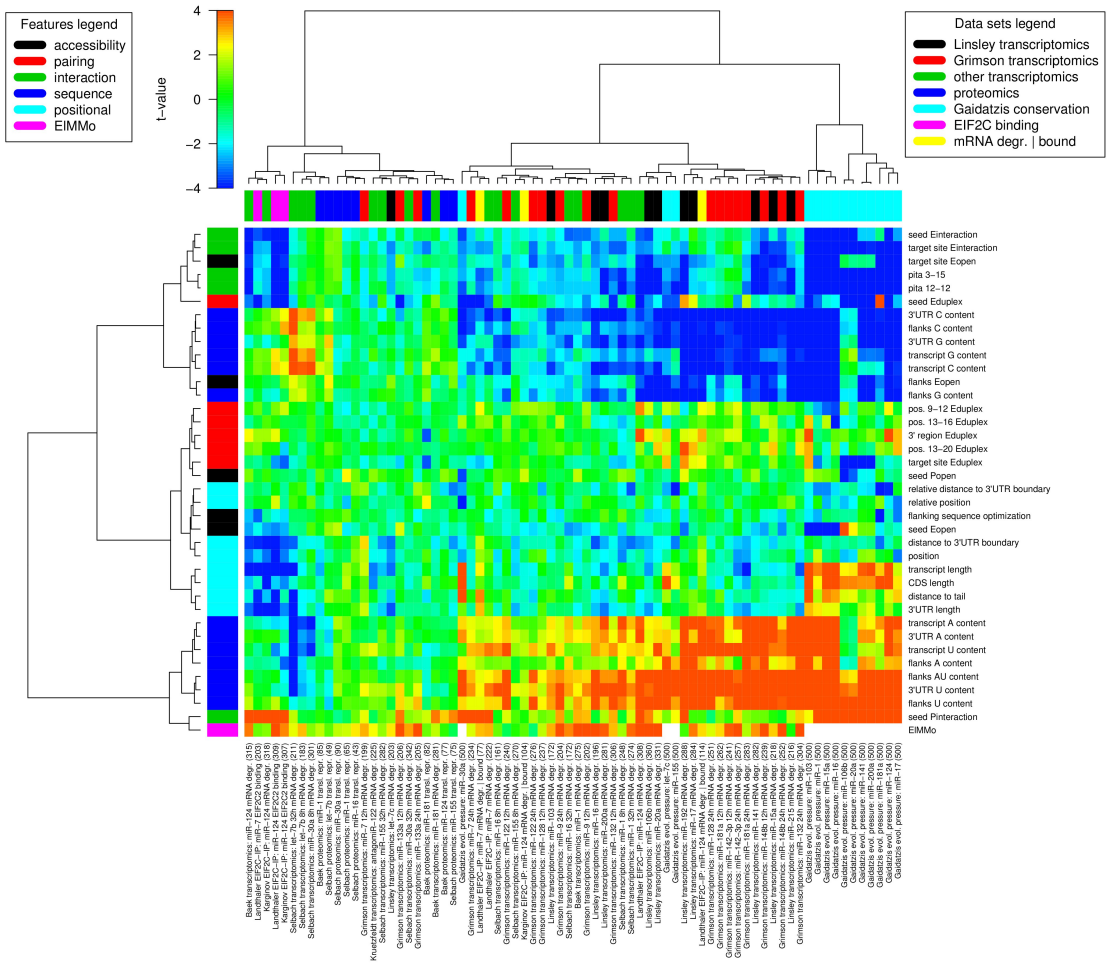


Figure 2: Predictive power of different features of putative miRNA target sites (rows) in predicting functional sites across the 74 data sets (columns). The data sets covered transcriptomics and proteomics measurements after miRNA transfection, transcriptomics measurements after miRNA knock-down, profiling of mRNAs bound to *EIF2C*/miRNA complexes, and target prediction based on comparative genomics. The column label indicates the source of the data set, the miRNA that was perturbed and the type of measurement that was performed, and the total number of sites involved in the analysis (positives + negatives). The heat-map shows the t-values comparing the distributions of feature values in functional vs non-functional 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. Supplementary Figure 17 in which the same data sets are sorted by the GC-content of the miRNA shows that the target site properties are not miRNA-specific.

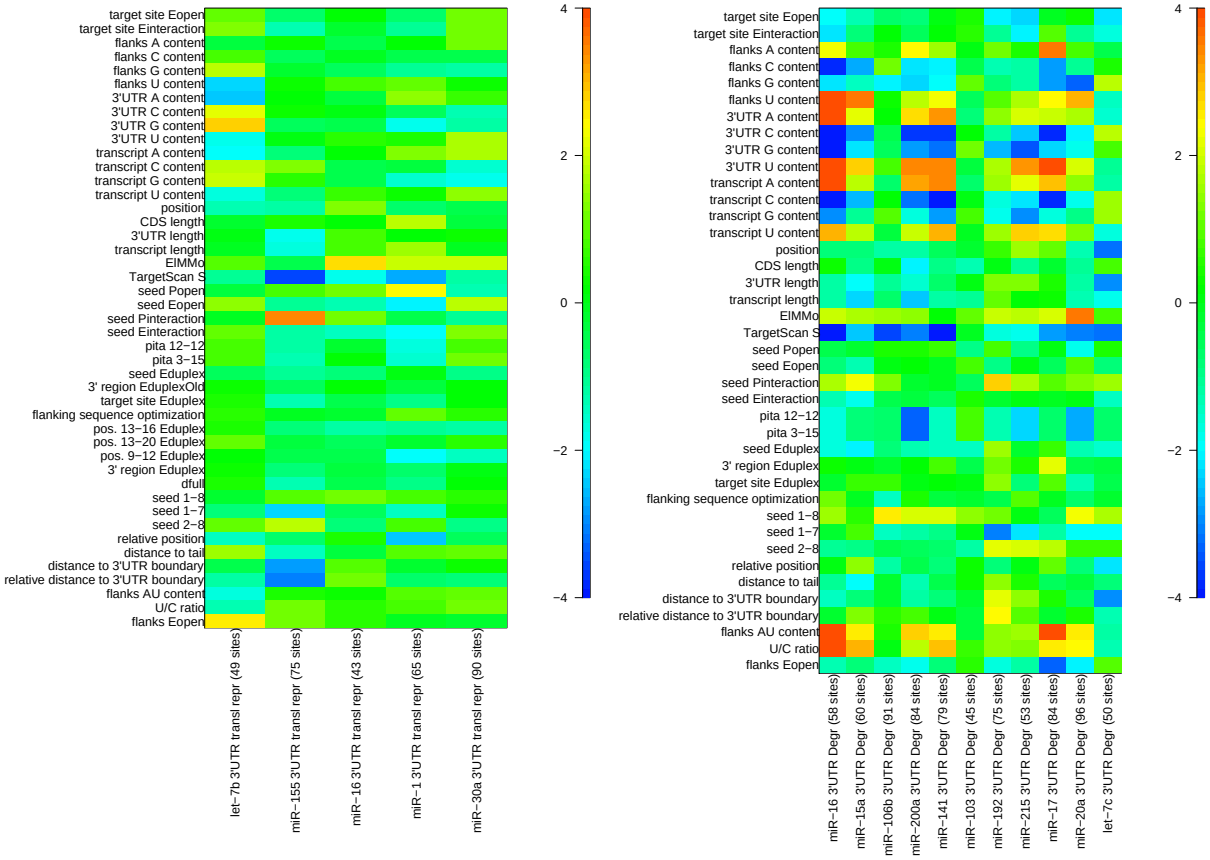


Figure 3: The smaller sample size in the proteomics miRNA transfection experiments cannot, on its own, explain the lack of predictive power that the features that we considered have for the proteomics data. Left panel: detail of the Supplementary Figure 2, showing the predictive power of different features on the proteomics experiments of Selbach et al. (2008). The shot-gun proteomics approach used by the authors (as well as by Baek et al. (2008)) makes it possible to quantify the change in concentration of 2000–3000 proteins following the transfection of a miRNA. Except for a few exceptions, most of the features we examined are not predictive of the functionality of miRNA binding sites in this series of experiments. Right panel: we replicated the feature analysis shown on Supplementary Figure 2 using only 2000 randomly selected genes from the miRNA transfection experiments analyzed with microarrays by Linsley et al. (2007). We then determined the predictive power of different features of the miRNA binding sites using the same selection criteria as for the proteomics datasets, i.e. comparing the 75 most down-regulated mRNAs to the 75 least regulated mRNAs. Despite the reduction of the sample size by a factor $3 - 7.5$, the predictive power of most sequence features as well as of some structure features is still detectable in most experiments. Therefore, the sample size cannot explain on its own why none of features we study appear to be predictive of the miRNA binding sites identified by the proteomics experiments.

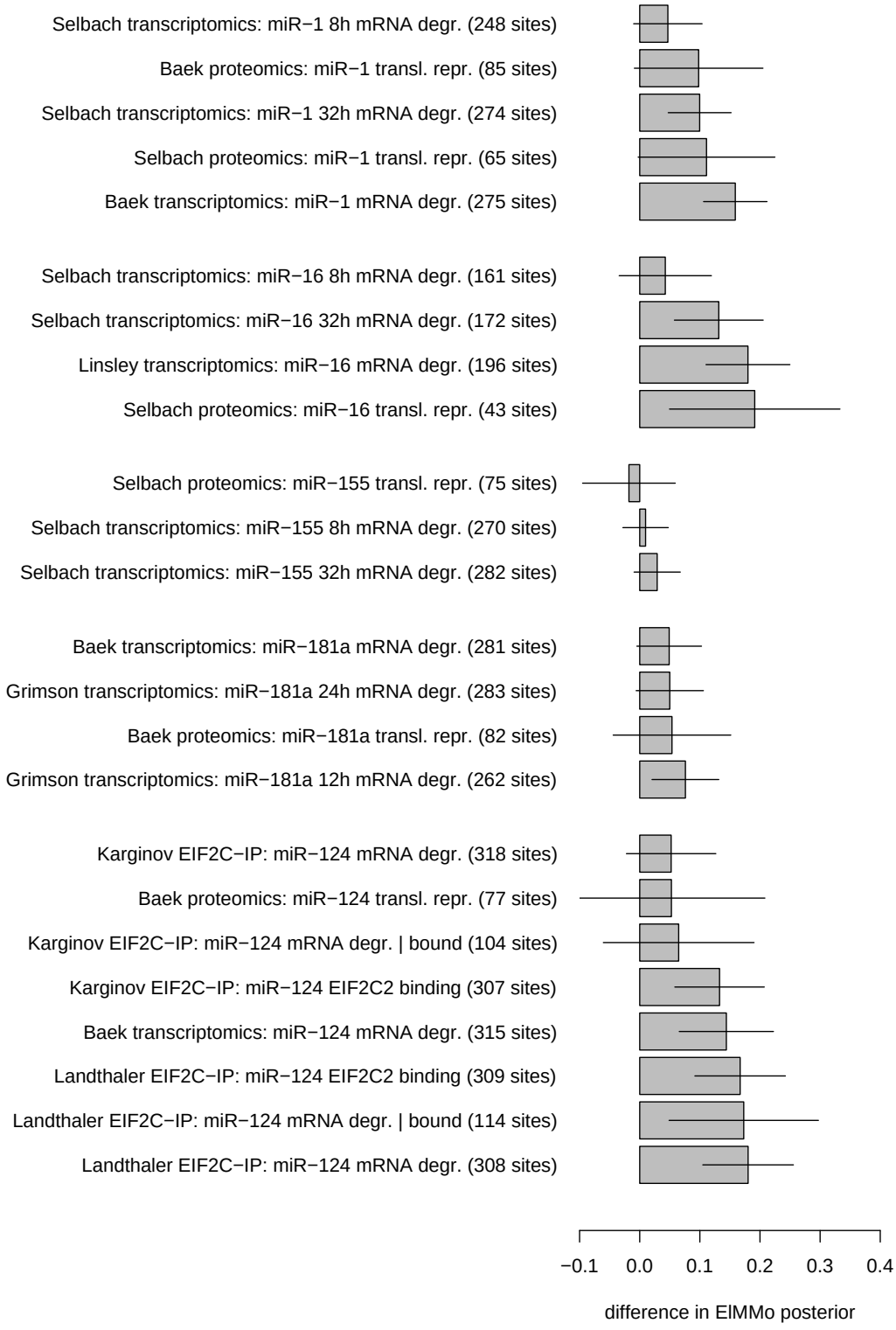


Figure 4: Difference between the average ElMMo posterior of functional vs non-functional miRNA target sites in different experiments.

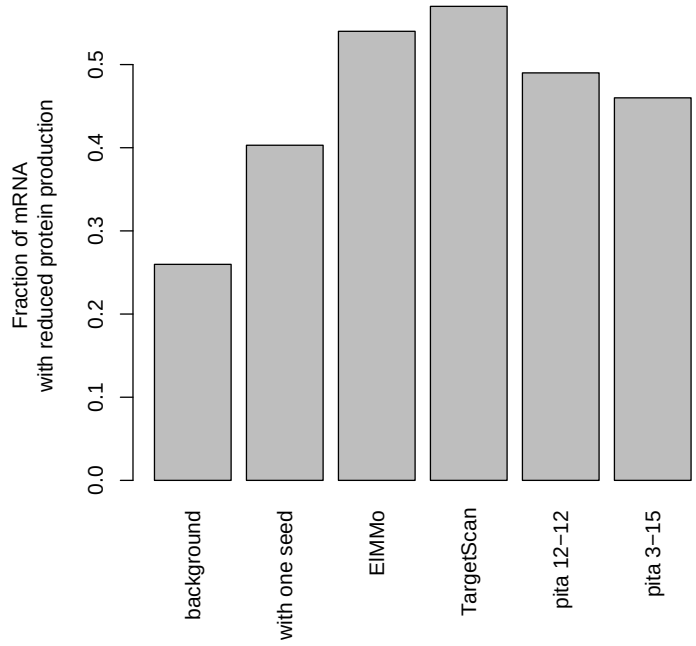


Figure 5: Fraction of the mRNAs obtained by applying a given "prediction" method that have reduced protein production according to the pSILAC experiments of Selbach et al. (2008).

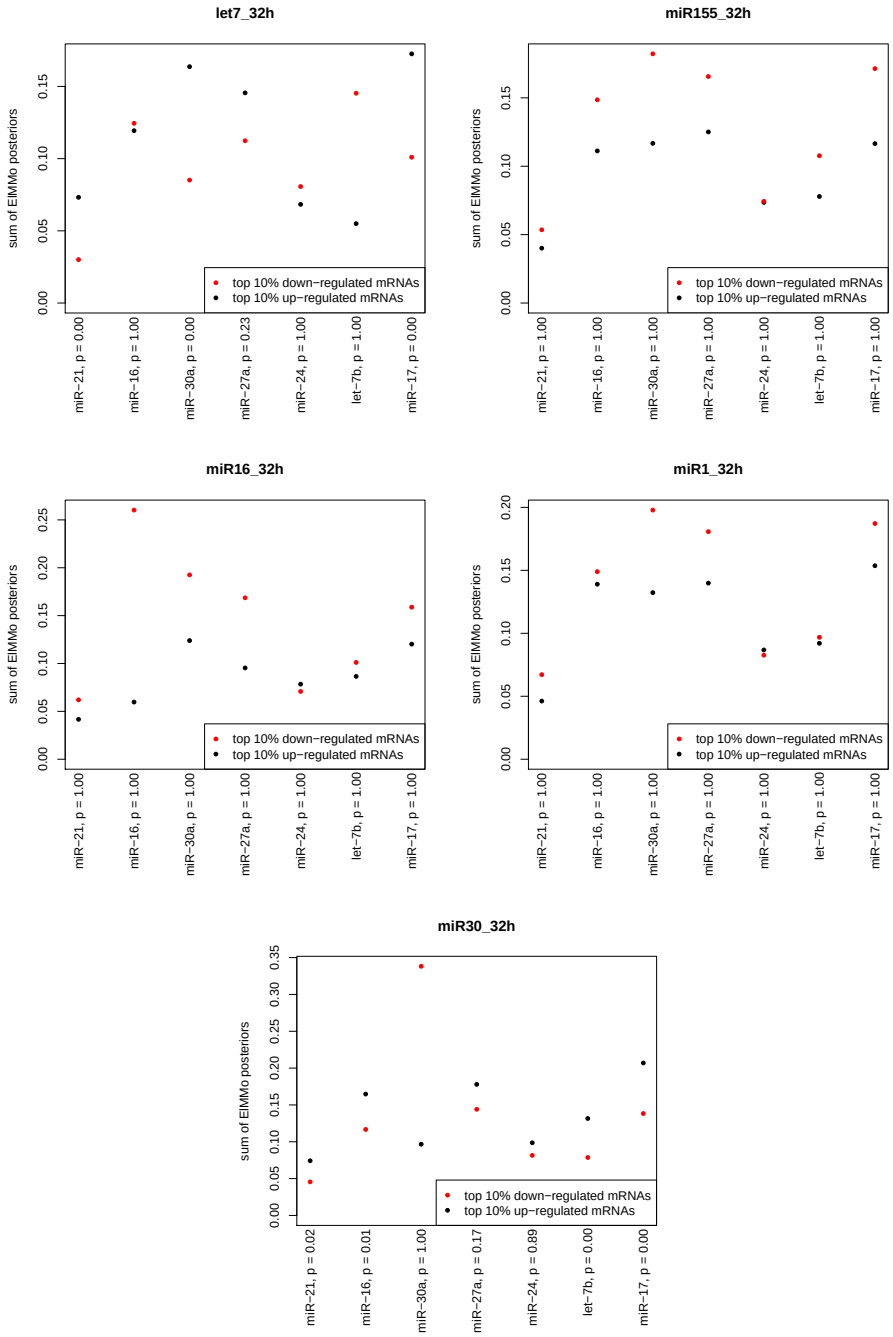


Figure 6: Expected number of evolutionarily selected binding sites for the 7 most abundant miRNAs in HeLa cells (Landgraf et al., 2007) in the 10% most up-regulated and down-regulated transcripts in individual transfection experiments of Selbach et al. (2008). The expected number of sites were computed by summing the ElMMo posteriors over all putative binding sites for a miRNA within every 3'UTR. Each panel represents one transfection experiment, where the transfected miRNA is indicated in the title. The expected number of binding sites in up- and down-regulated transcripts where compared using Wilcoxon's ranks sum test. The corresponding p-values were computed under the alternative hypothesis that up-regulated transcripts harbor more miRNA binding sites under evolutionary pressure than down-regulated transcripts and were corrected for multiple testing using the Bonferroni method.

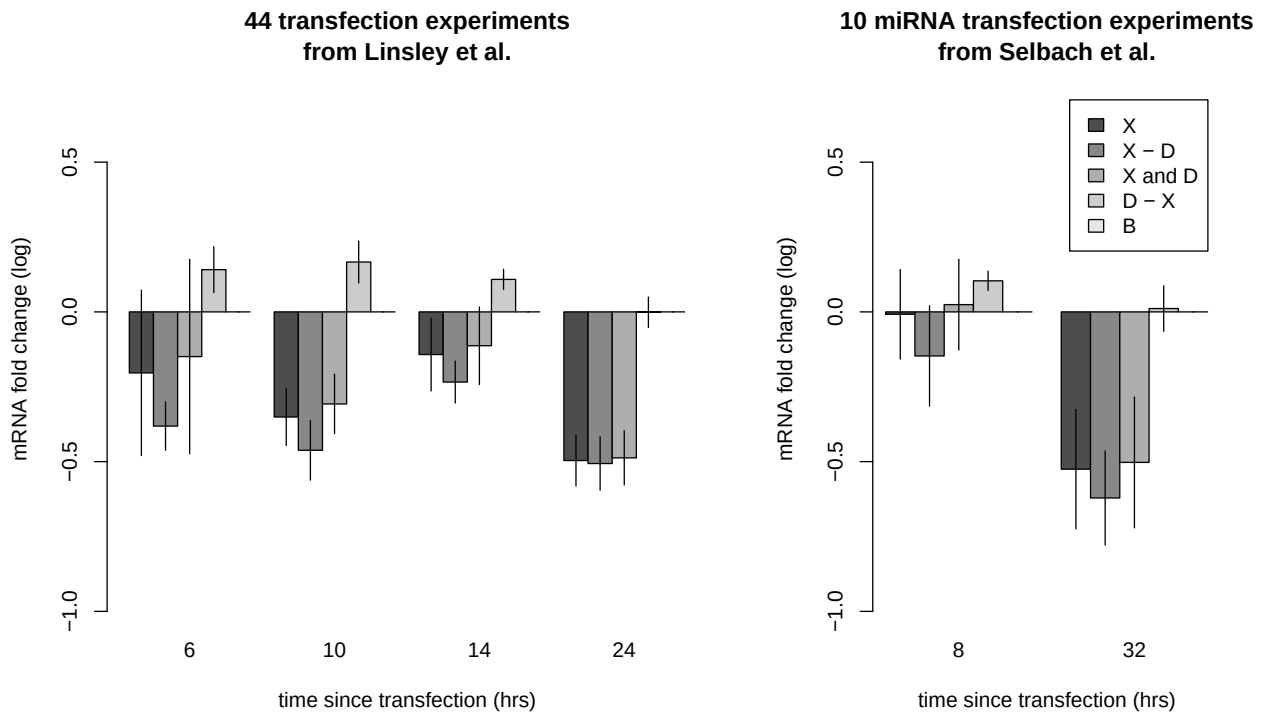


Figure 7: The competition between endogenous miRNAs and the transfected miRNA is transient in time. The y-axis shows average the log fold change of mRNAs carrying seed matches to the transfected miRNA in their 3' UTR (X), seed matches to the transfected miRNA but not to the most expressed endogenous miRNAs (X - D), seed matches to both the transfected miRNA and the top expressed endogenous miRNAs (X and D), seed matches to the most expressed endogenous miRNAs but not the transfected miRNA (D - X), and seed matches to neither the transfected miRNA nor the endogenous miRNAs (B). The error bars show the 95% confidence interval on the mean after averaging over all miRNA transfection experiments performed at the same time point. Left: re-analysis of 44 microarray experiments performed 6, 10, 14 and 24 hours after miRNA / siRNA transfection in HCT116 Dicer -/- cells (Linsley et al., 2007). Right: re-analysis of 10 microarray experiments performed 8 and 24h after miRNA transfection in HeLa cells (Selbach et al., 2008).

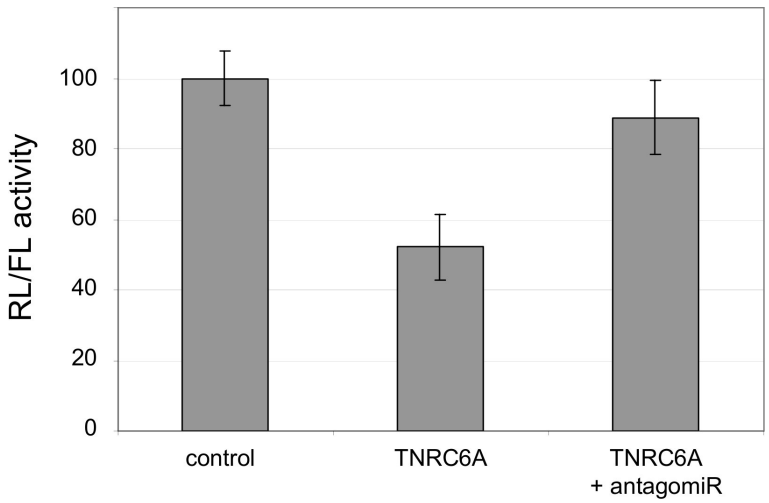


Figure 8: Luciferase reporter assay confirming that *TNRC6A* (also known as *GW182*) is a direct target of the endogenously expressed miR-30a in HeLa cells. The *TNRC6A* 3'UTR was cloned downstream of the coding region of the *Renilla* luciferase (RL) and the vector system subsequently transfected into HeLa cells, either alone (*TNRC6A*) or together with the miR-30a antisense inhibitor (*TNRC6A* + antagomiR). The y-axis shows the change in *Renilla* luciferase activity normalized to the firefly luciferase (FL, control).

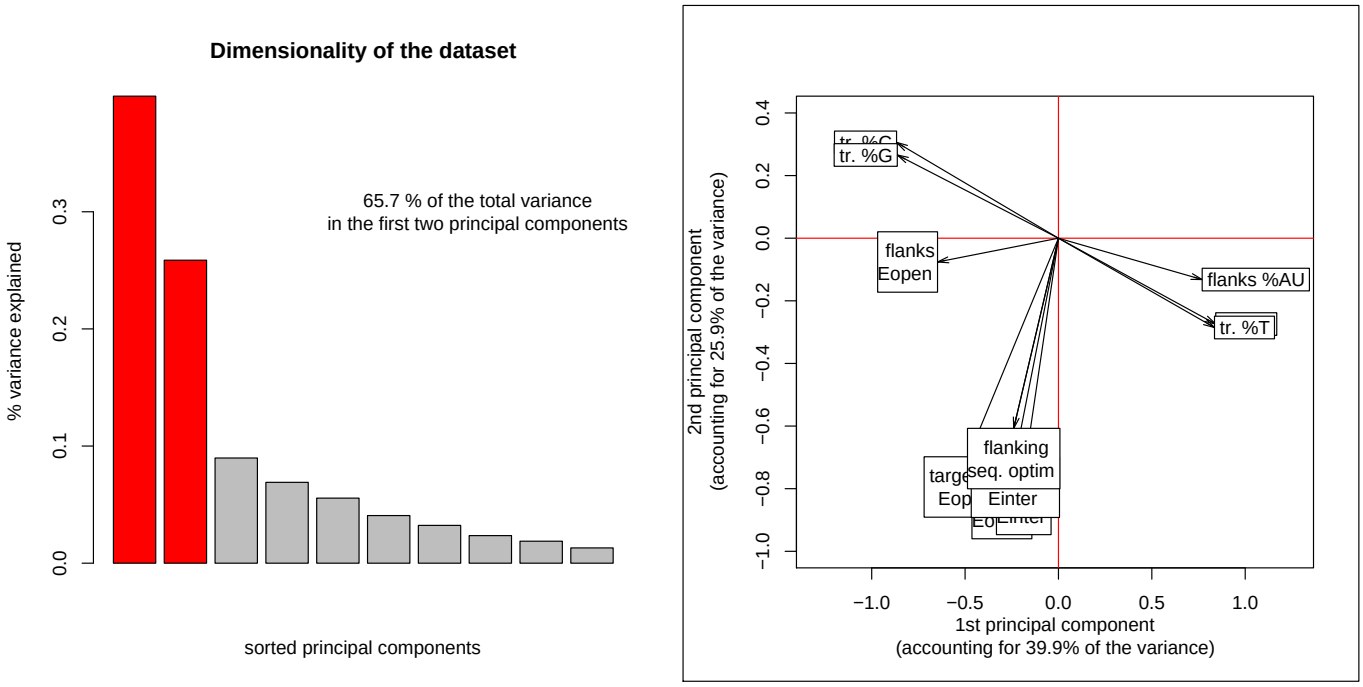


Figure 9: Principal component analysis of a subset of features computed over 5964 miRNA binding sites (positives and negatives) from the comparative genomics data set. Although Figure 2 shows that, when considered independently, both structure as well as sequence features are predictive of miRNA target site functionality, it does not show to what extent these features are redundant. To determine this, we collected all 5964 miRNA binding sites from the comparative genomics dataset (comprising positives as well as negatives) and considered the following set of features: transcript A, G, C and U content, flanks AU content, the seed, target site and flanks Eopen, flanking sequence optimization, seed and target site Einteraction. We then centered and rescaled this subset of features and determined how many principal components are needed to describe this subset of features. The first two principal components accounted for 65.7% of the total variance, with the third component and next components accounting for a substantially smaller amount of the variance compared to the first two principal components (left panel). We then projected the subset of features onto the plane spanned by the first two principal components and determined that sequence features clustered well with the first principal component, while all structure features except "flanks Eopen" clustered together with the second principal component (right panel). This suggests that, except for "flanks Eopen" which correlates with the G and C content, sequence and structure features are not redundant and characterize miRNA binding sites in a complementary way. Performing the same analysis on the smaller transcriptomics and proteomics datasets yielded the same results.

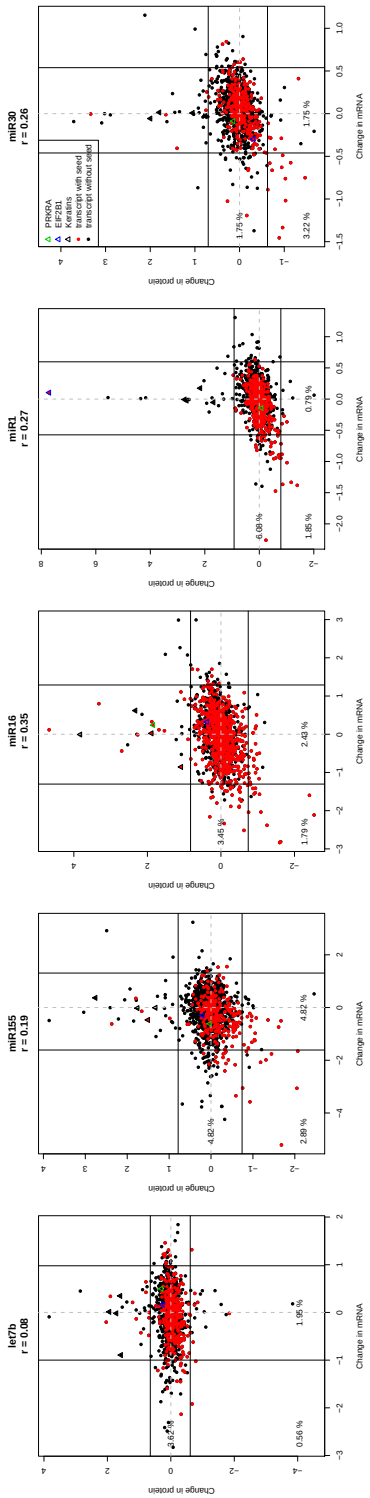


Figure 10: Correlation between change in protein and mRNA levels in the let-7, miR-155, miR-16, miR-1 and miR-30a pSILAC experiments of Selbach et al. (2008). The x-axis shows the log2 fold change in expression between miRNA-transfected to mock-transfected HeLa cells. The black lines indicate the cut-offs in mRNA and protein level fold change beyond which we consider the mRNA or the protein differentially expressed. Red and black dots respectively represent transcripts that carry at least one match or do not carry any match to the seed of the transfected miRNA. The three percentages respectively indicate the proportion of transcripts carrying at least on miRNA seed match that are down-regulated at the protein level only, at the mRNA level only, or both at the levels of the protein and mRNA. r is the Pearson correlation coefficient between the change in protein and mRNA levels.

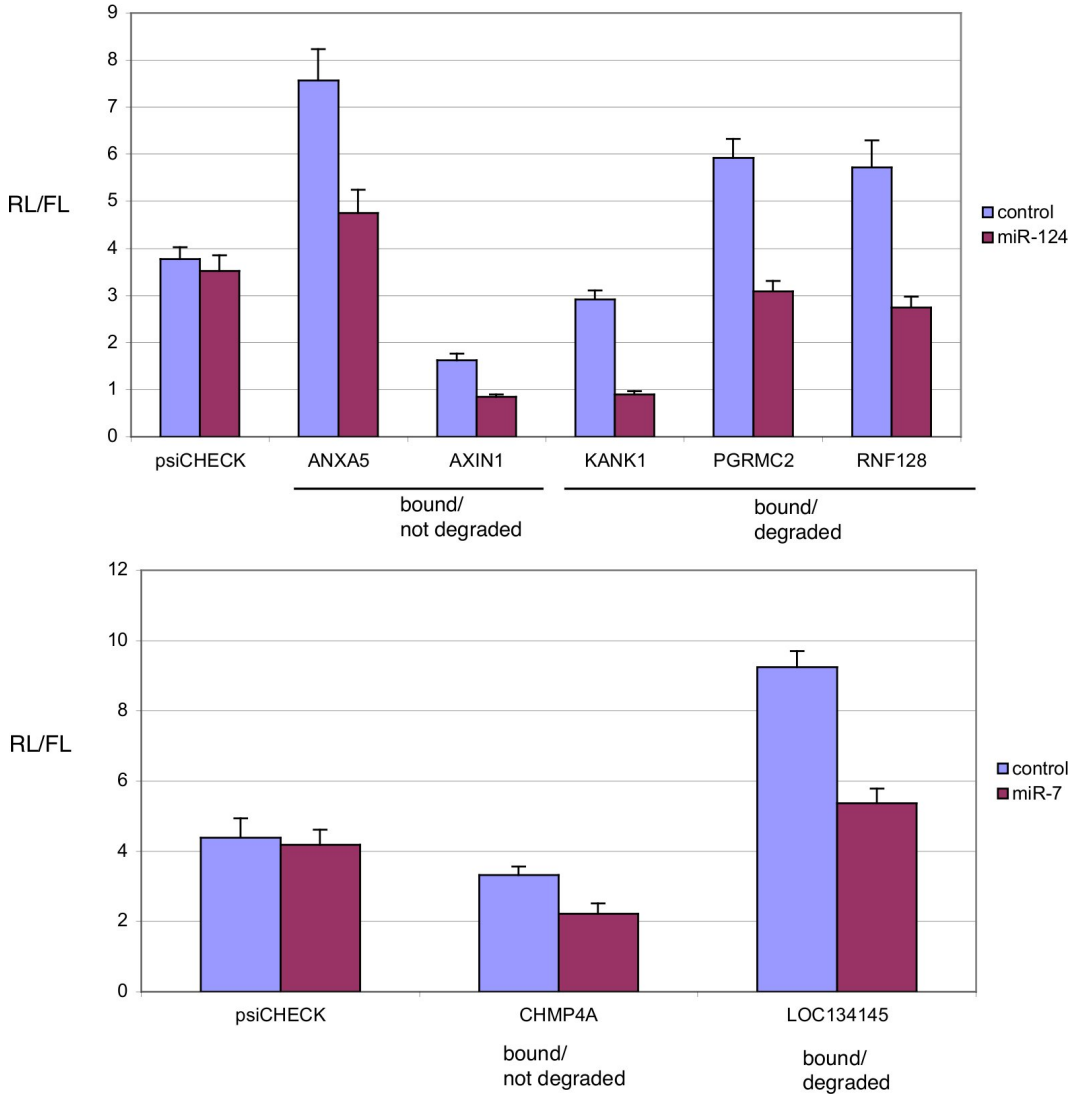


Figure 11: miR-124 and miR-7-mediated repression of 3'UTRs fused to luciferase reporter genes. psiCHECK reporter constructs were generated by fusing the full-length 3'UTRs of the genes indicated to the *Renilla* Luciferase. Dual Luciferase activity from HEK293 cells cotransfected with each reporter psiCHECK construct and miR-124 or miR-7 duplex was compared to cotransfection of each reporter construct with control RNA duplex. Transfections of parental psiCHECK vector without inserted 3'UTR (psiCHECK) is shown. *Renilla* Luciferase versus firefly luciferase activities are indicated. Error bars represent standard deviation computed over 10 replicates.

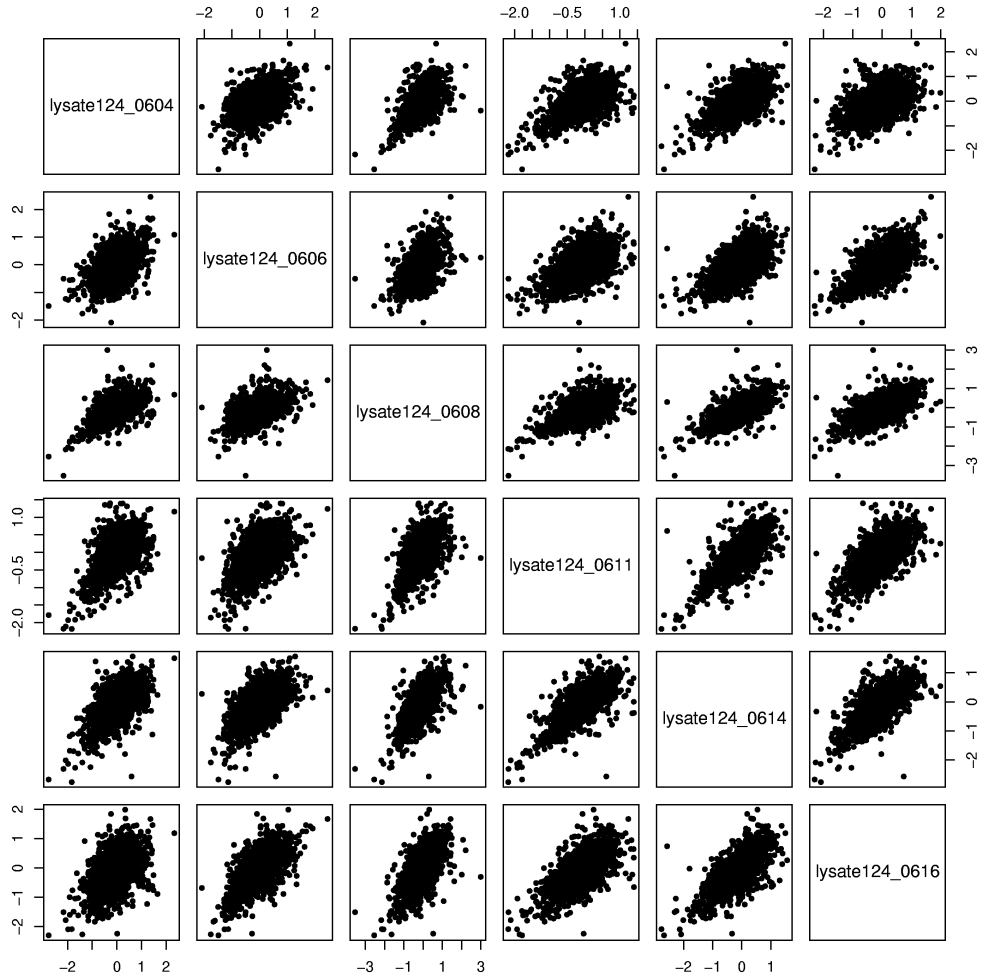


Figure 12: Correlation between the extent of mRNA degradation following miR-124 transfection in the 6 biological replicates of the transcripts of the Karginov et al. *EIF2C2*-IP dataset. The axes show log10 fold changes in pairs of replicates. The Pearson correlation coefficient between log10 fold changes of replicates ranges from 0.44 to 0.76 depending on the pair of experiments being considered.

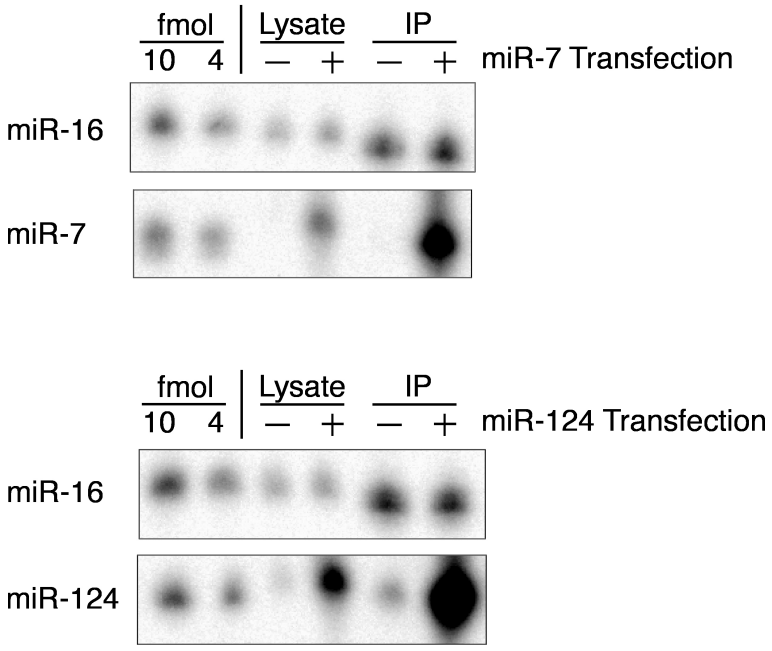
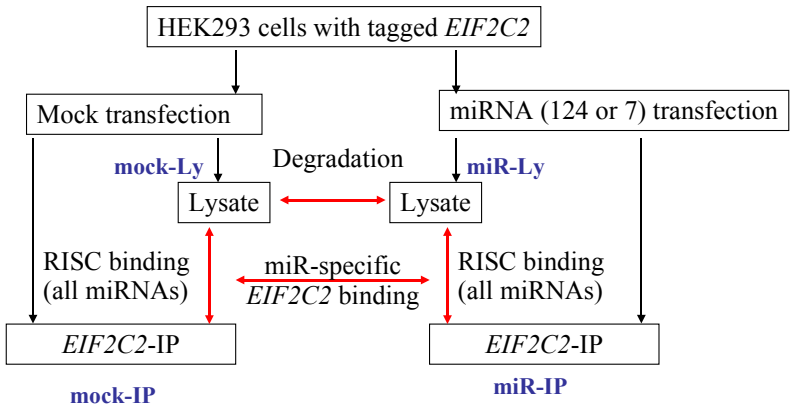


Figure 13: miRNA transfection and immunoprecipitation. Cells stably expressing FLAG/HA-EIF2C2 were mock-transfected (-) and transfected with a miR-7/miR-7* and miR-124/miR-124* duplex (+), respectively. 15 hours after transfection cells were lysed and the epitope-tagged protein was immunoprecipitated from cytoplasmic extracts with FLAG-antibody. RNA was extracted from the cleared cell lysate and the immunoprecipitate (IP). 15 μg total cellular RNA and one fifth of IPed RNA was separated on a 12% polyacrylamide gel, blotted, and probed for miR-16, miR-7, and miR-124, respectively. 10 and 4 femtomole (fmol) of synthetic miR-16, miR-7, and miR-124, were loaded as standards.



Degradation:	miR-Ly / mock-Ly
EIF2C2-binding miRNA:	miR-IP / miR-Ly
EIF2C2-binding mock:	mock-IP / mock-Ly
miR-specific *EIF2C2*-binding:	(miR-IP / miR-Ly) / (mock-IP / mock-Ly)

Figure 14: Sketch of the computation of the binding and degradation measures: *EIF2C2*-binding in miRNA transfection is given by the ratio of transcript levels in the immunoprecipitate and in the lysate of miR-transfected cells (miR-IP/miR-Ly); *EIF2C2*-binding in mock-transfection is given by the ratio of transcript levels in the immunoprecipitate and in the lysate of mock-transfected cells (mock-IP/mock-Ly); miR-specific *EIF2C2*-binding is given by the ratio of the previous two ratios, (miR-IP/miR-Ly)/(mock-IP/mock-Ly).

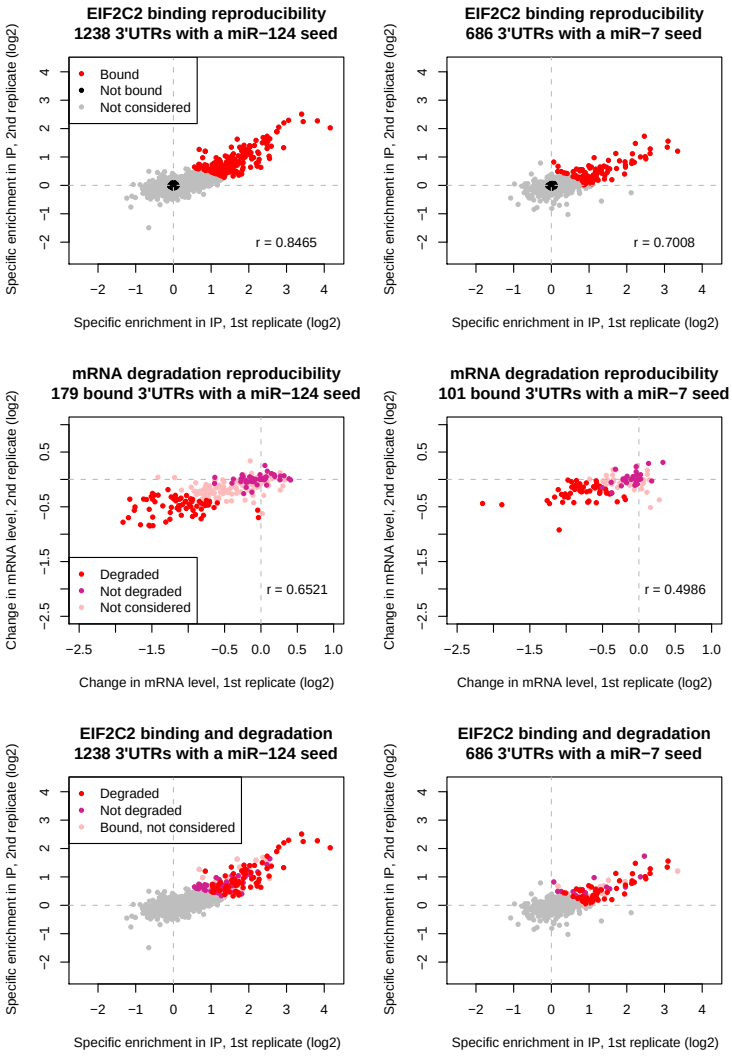


Figure 15: Selection of positive and negative examples for *EIF2C2* binding and mRNA degradation upon miR-124 (left) and miR-7 (right) transfection. Upper panels show the correlation between *EIF2C2* binding measures in the replicate experiments. The transcripts marked with red were considered "bound" and those marked in black "not bound". The procedure for this selection is described in the supplementary text. Middle panels show the correlation between degradation of bound transcripts (in red in the upper panels) in replicate experiments. Transcripts marked in red were considered "bound and degraded", those in violet "bound but not degraded". Lower panels reproduce the upper panels, except that transcripts that were considered "bound" are further shown in the color that indicates whether they were or not also considered degraded. These figures show that the degree of degradation is not simply proportional to the degree of *EIF2C2* binding.

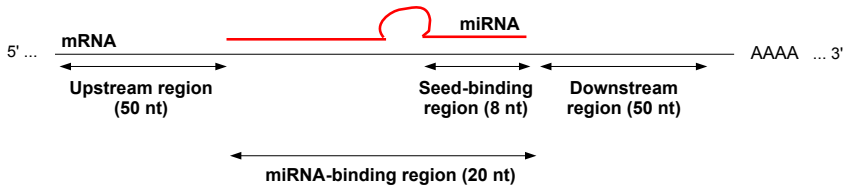


Figure 16: Sketch of the transcript regions used in the computation of structural and sequence features.

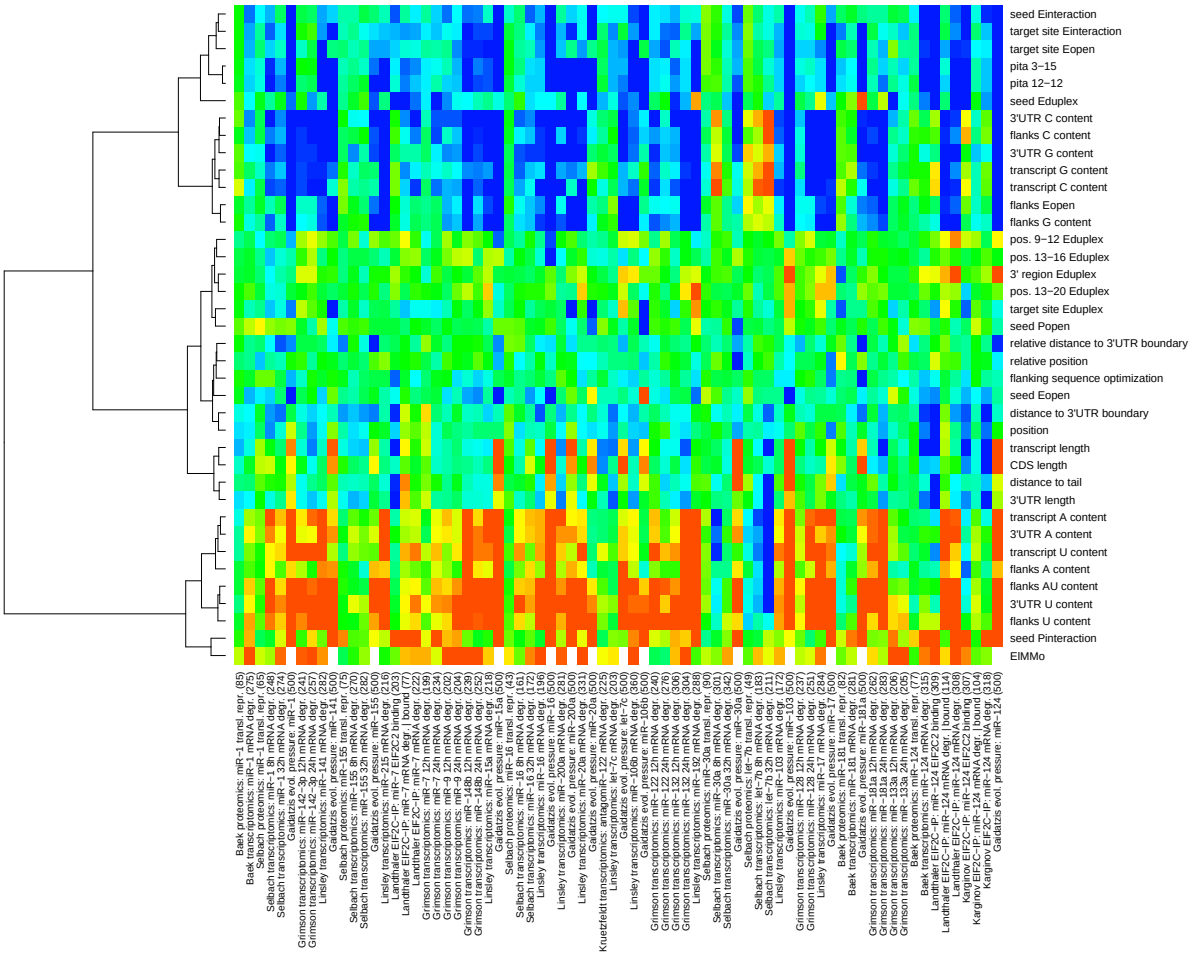


Figure 17: The features predictive of miRNA targeting are not determined by the GC content of the mature miRNA. The present figure shows a heat map similar to ones shown on Figure 1 and Supplementary Figure 2, but in which we reordered the columns (data sets) according to the GC content of the transfected miRNA. The left-most columns correspond to GC-poor miRNAs while the right-most columns feature data sets involving GC-rich miRNAs.