A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets

Mohsen Khorshid[1,2], Jean Hausser[1,2], Mihaela Zavolan[1,2] & Erik van Nimwegen[1,2]

We introduce a biophysical model of miRNA-target interaction and infer its parameters from Argonaute 2 cross-linking and immunoprecipitation data. We show that a substantial fraction of human miRNA target sites are noncanonical and that predicted target-site affinity correlates well with the extent of target destabilization. Our model provides a rigorous biophysical approach to miRNA target identification beyond *ad hoc* miRNA seed–based methods.

© 2013 Nature America, Inc. All rights reserved.

MicroRNAs (miRNAs) are a large class of regulators of gene expression that post-transcriptionally modulate the stability of mRNA targets and their rate of translation into proteins. Although in mammals, 7 or 8 nucleotides of perfect complementarity between the miRNA 5′ end and the target mRNA are frequently sufficient to elicit a response (typically measured in terms of mRNA degradation[1]), many such 'miRNA seed'–matching sites have no apparent effect. Thus, current target prediction methods additionally make use of conservation and sequence context information to reduce false positive predictions[2,3]. 'Noncanonical' sites, which are not perfectly complementary to the miRNA seed region yet are effective in downregulating gene expression, have also been described[4,5]. However, these are considered rare, and the currently most accurate prediction methods do not attempt to identify them.

Recently developed methods for Argonaute protein cross-linking and immunoprecipitation (Ago-CLIP)[6] enable experimental identification of miRNA binding sites transcriptome wide. Although this provides the opportunity to investigate in detail the principles and determinants of miRNA-mRNA target interaction, Ago-CLIP on its own does not identify which miRNA guided Ago to each binding site or the structure of the miRNA–target site hybrid. Here we introduce a rigorous biophysical model of miRNA-target interaction and infer its energy parameters from Ago-CLIP data. The model (which we called MIRZA; see Online Methods) includes parameters associated with base pairs and loops and specific miRNA position–dependent energy parameters that reflect the constraints imposed by the Argonaute protein on miRNA-mRNA interaction. The process by which MIRZA calculates the energy of a possible miRNA-mRNA target hybrid in terms of its 27 energy parameters is illustrated in **Figure 1a**.

Given a set of parameters, MIRZA predicts the frequencies with which RNA-induced silencing complexes (RISCs) bind to different mRNA fragments in the mRNA pool. We infer MIRZA's parameters by maximizing the binding probabilities of the mRNA fragments observed in an Ago2-CLIP sample (Online Methods). This involves calculating a 'target quality' $R(m|\mu)$ that quantifies the total affinity of each miRNA μ for each fragment m. Specifically, $R(m|\mu)$ corresponds to the enrichment of fragment m among target sites bound by miRNA μ, relative to m's abundance in the mRNA pool. Calculating $R(m|\mu)$ involves summing over all possible hybrid structures that m can form with μ. The fraction of time that fragment m is bound by a RISC loaded with miRNA μ is proportional to the 'target frequency' $R(m|\mu)\pi_\mu$, which depends on the fractions π_μ of mRNA-bound RISCs loaded with miRNA μ. These fractions, which we call miRNA priors, are inferred for each given CLIP data set. The overall probability of immunoprecipitating fragment m relative to its background frequency is then given by $R(m) = \Sigma_\mu R(m|\mu)\pi_\mu$, and the likelihood of the entire data set by the product $R(D) = \prod_i R(m_i)$ over all observed fragments m_i.

We first tested the procedure on synthetic data sets containing seed-matching sites and 3′ compensatory sites similar to those previously described[7]. MIRZA successfully inferred the energy parameters that were used in generating these synthetic data sets and perfectly predicted which miRNA was associated with each site (**Supplementary Note**). To infer the energy parameters of real miRNA-target interactions from Ago2-CLIP data, we used 2,988 mRNA regions that were reproducibly cross-linked in at least three of four Ago2-CLIP data sets from ref. 8 (**Supplementary Table 1**) and included all miRNAs that were expressed in the HEK293 cells in which the experiments were performed (Online Methods). Ago2 was found to preferentially cross-link to nucleotides located in the center of the hybrid between the target site and the miRNA. These nucleotides are easily identifiable through diagnostic mutations that are introduced during cDNA preparation and are used to pinpoint the miRNA binding sites with very high resolution[8]. Thus, to generate our input set of miRNA binding sites, we extracted 51-nucleotide-long regions centered on the position with the highest number of cross-link diagnostic mutations. We performed 100 parameter-optimization runs starting with randomly chosen initial values for all parameters.

Different optimization runs yielded highly reproducible parameter sets (**Fig. 1b** and **Supplementary Note**). Consistent with the known importance of the seed region, positions 2–7 have the largest positive contribution to the energy (parameters E_2–E_7

[1]Biozentrum, University of Basel, Basel, Switzerland. [2]Swiss Institute of Bioinformatics, Basel, Switzerland. Correspondence should be addressed to M.Z. (mihaela.zavolan@unibas.ch) or E.v.N. (erik.vannimwegen@unibas.ch).

RECEIVED 28 MAY 2012; ACCEPTED 16 DECEMBER 2012; PUBLISHED ONLINE 20 JANUARY 2013; DOI:10.1038/NMETH.2341

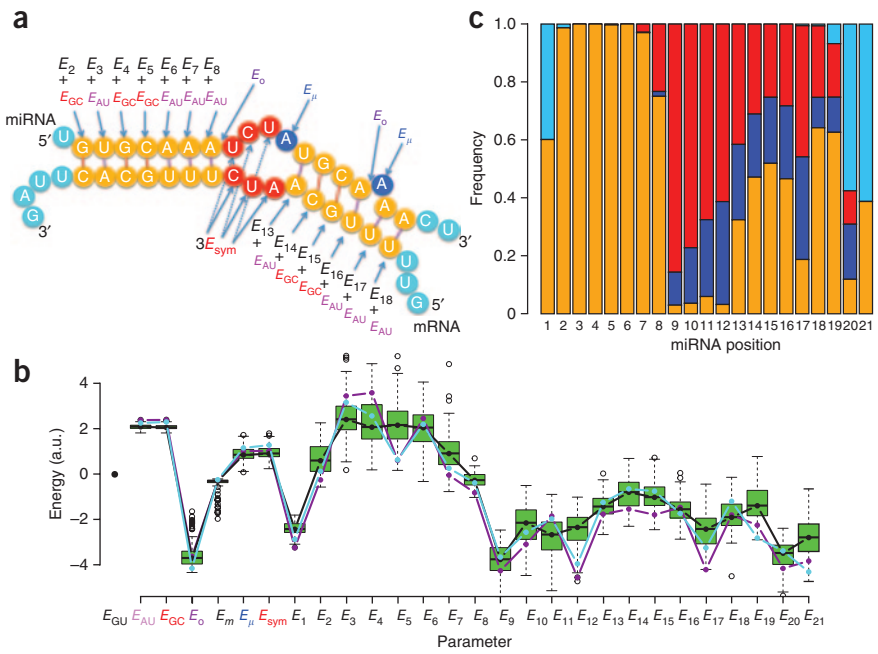

Figure 1 | A biophysical model of miRNA-target interaction. (**a**) Sketch of a miRNA-mRNA hybrid illustrating the way MIRZA assigns a binding energy to the interaction. Nucleotides involved in base-pairing are indicated in orange, symmetric loops in red, bulges in the miRNA in blue and dangling ends in cyan. Arrows point from the independent energy terms (defined in Online Methods) to the corresponding structural elements (base pairs, loop openings and extensions). (**b**) Summary of energy parameters inferred from 100 independent optimization runs on the Ago2-CLIP data. Green boxes show interquartile ranges, the 5th and 95th percentiles are indicated by whiskers, and black dots indicate median values of fitted parameters across the runs. The sets of parameters that yielded the highest and second-highest probability are shown as purple and cyan dots, respectively. a.u., arbitrary units. (**c**) Summary of the predicted hybrid structures; colors indicate the fraction of hybrids in which a given nucleotide is involved in a base pair (orange), symmetric loop (red), bulge (blue) or dangling end (cyan).

in **Fig. 1b**), followed by positions 13–16 (E_{13}–E_{16}) and 18 and 19 (E_{18} and E_{19}). In contrast, hybridization of position 9 (E_9) is strongly disfavored, as is opening a loop (E_o). Once a loop is opened, symmetric loops (E_{sym}) and bulges in the miRNA (E_μ) are clearly favored over bulges in the mRNA (E_m).

With the fitted parameters we can predict which miRNA μ is most likely to bind each fragment m, as well as the structure of the most likely hybrid between m and μ. **Figure 1c** statistically summarizes the structures of these predicted hybrids. Notably, even though no specific knowledge about miRNA-target interactions went into the inference of its parameters, and in contrast to general models of RNA-RNA interaction applied to the same data (**Supplementary Note**), our model captures several known structural features of miRNA-target interaction such as the predominant binding of the nucleotides in the seed region, the less frequent binding of position 1, and the possibility of compensatory base-pairing at the miRNA's 3′ end. MIRZA makes the specific prediction that nucleotides 14–16 of the miRNA are base-paired with the target roughly 50% of the time and that positions 18 and 19 are bound even more than 60% of the time.

For more than 26% of the most enriched, reproducibly cross-linked sites, the most likely hybrid is noncanonical (**Supplementary Note**). This is noteworthy because functional noncanonical sites are thought to be rare, and the more accu-

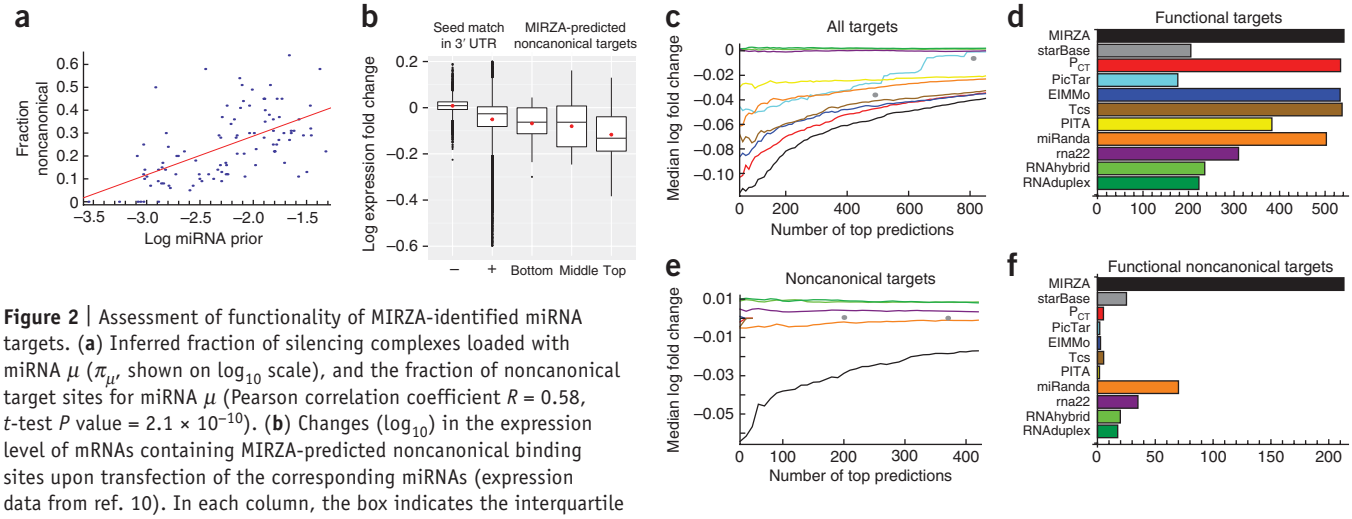

Figure 2 | Assessment of functionality of MIRZA-identified miRNA targets. (**a**) Inferred fraction of silencing complexes loaded with miRNA μ (π_μ, shown on $\log_{10}$ scale), and the fraction of noncanonical target sites for miRNA μ (Pearson correlation coefficient $R = 0.58$, t-test P value = 2.1×10^{-10}). (**b**) Changes ($\log_{10}$) in the expression level of mRNAs containing MIRZA-predicted noncanonical binding sites upon transfection of the corresponding miRNAs (expression data from ref. 10). In each column, the box indicates the interquartile range, the black line the median, the red dot the mean, and the whiskers the 5th and 95th percentiles. The first two columns correspond to transcripts without and with seed matches for the transfected miRNA, and the last three columns correspond to transcripts containing a noncanonical, MIRZA-predicted site with a target quality score among the lowest, middle and highest 33%. (**c**) Median log fold change of targets predicted by MIRZA (black), TargetScan P_{CT}[2] (red), PicTar[15] (cyan), EIMMo[3] (dark blue), TargetScan context+[16] (brown), PITA[17] (yellow), miRanda[18] (orange), rna22 (ref. 19) (violet), RNAhybrid[20] (light green) and RNAduplex[21] (dark green), averaged over 38 transfection experiments from five studies[10–14]. The gray dots show fold changes of targets obtained by intersecting Ago-CLIP with computationally predicted sites (starBase database[22]). (**d**) Estimated total number of functional targets predicted by different methods averaged over all transfection experiments. Colors as in **c**. (**e**,**f**) Plots as in **c** and **d**, respectively, but considering only noncanonical targets.

© 2013 Nature America, Inc. All rights reserved.

rate current target prediction methods focus solely on canonical sites. However, recent experimental studies[9] hinted that noncanonical sites may be more prevalent, particularly those in which an mRNA nucleotide is bulged out between positions 5 and 6. Applying MIRZA to the data from ref. 9, we indeed find that, depending on the sample, 9%–20% of the predicted miR-124 sites correspond to this particular noncanonical site (**Supplementary Table 2**). MIRZA, however, predicts several other types of noncanonical sites, such as contiguous pairing of only nucleotides 2–6, in all CLIP data sets.

MIRZA further infers that the fraction of noncanonical sites is higher for miRNAs with the highest abundance in RISCs—that is, for those with high prior π_μ—and that the fraction of noncanonical sites can be as high as 60% (**Fig. 2a**). The inferred abundance π_μ correlates significantly with the expression level of the miRNA, suggesting that the target spectrum of a miRNA depends crucially on its expression level: miRNAs with low expression target mainly high-affinity canonical sites, whereas miRNAs with high expression target those sites and also large numbers of noncanonical sites, which, on average, have lower affinity (**Supplementary Note**).

Gene expression analysis shows that the noncanonical sites inferred from the CLIP data are functional, inducing a significant downregulation of host transcripts upon miRNA transfection (**Fig. 2**, z values of –4.44, –4.89 and –6.01 for the three categories of noncanonical sites; and Online Methods). Although sites with higher predicted target quality show stronger downregulation, even transcripts containing the weakest noncanonical sites show stronger downregulation than transcripts that simply carry seed matches (**Fig. 2b**). That noncanonical sites show significantly more evolutionary conservation than is expected by chance ($P = 0.0048$; **Supplementary Note**) is further indication of their functionality.

To compare the accuracy of the target sites identified by MIRZA in Ago2-CLIP data with those of miRNA target prediction methods, we analyzed 38 transfection experiments involving 26 different miRNAs[10–14], comparing the miRNA-induced fold changes of transcripts predicted by these methods (Online Methods and **Supplementary Note**). To assess the ability of a method to identify the most strongly downregulated targets, we sorted its predicted targets by their scores and calculated the median fold change of the top n targets as a function of n (**Fig. 2c**). To assess the total number of functional targets predicted by a method, we calculated how many more targets were downregulated compared to the number expected by chance (**Fig. 2d**). Although the relative performance of the different methods varies across data sets, MIRZA's predictions show the strongest downregulation on average (**Fig. 2c**) and for the large majority of individual data sets and miRNAs (**Supplementary Note**). Furthermore, in terms of the total number of functional targets that it predicts (**Fig. 2d** and **Supplementary Note**), MIRZA matches the best methods that use evolutionary conservation (TargetScan P_{CT} and ElMMo) or the context of the sites (TargetScan context and miRanda).

It is in the prediction of functional noncanonical targets that MIRZA's performance stands out (**Fig. 2e**,**f** and **Supplementary Note**). MIRZA identifies at least threefold more functional targets than any other method, and its targets undergo much stronger downregulation, which is strongly correlated with their MIRZA score (**Fig. 2e**). Moreover, this performance is consistent across all data sets and individual miRNAs (**Supplementary Note**). Finally, the partial overlap between the sites identified for some miRNAs by MIRZA and by algorithms based on conservation or context suggests that miRNA target prediction could be further improved by combining MIRZA's biophysical model with context and conservation information.

MIRZA is made available among the tools provided on our CLIPZ server (http://www.clipz.unibas.ch/).

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to N. Beerenwinkel and S. Bergmann for comments in the initial stages of this work. We are also thankful to A.R. Gruber and the other members of the Zavolan group for providing input and feedback on the algorithm and the manuscript, A. Crippa for help with the code distribution and P.J. Balwierz for help converting the LaTeX manuscript to Word. M.K. was supported by Swiss National Science Foundation ProDoc grant PDFMP3_123123 to M.Z. and E.v.N. The work was additionally supported by Swiss National Science Foundation grant 31003A_127307 to M.Z.

AUTHOR CONTRIBUTIONS

Conceived of and designed the experiments: E.v.N. and M.Z. Performed the experiments: M.K. and J.H. Analyzed the data: J.H., M.K., E.v.N. and M.Z. Wrote the paper: J.H., M.K., M.Z. and E.v.N.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2341.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bartel, D.P. *Cell* **136**, 215–233 (2009).
2. Friedman, R.C., Farh, K.K.H., Burge, C.B. & Bartel, D.P. *Genome Res.* **19**, 92–105 (2009).
3. Gaidatzis, D., van Nimwegen, E., Hausser, J. & Zavolan, M. *BMC Bioinformatics* **8**, 69 (2007).
4. Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K. & Slack, F.J. *Genes Dev.* **18**, 132–137 (2004).
5. Lal, A. *et al. Mol. Cell* **35**, 610–625 (2009).
6. Chi, S.W., Zang, J.B., Mele, A. & Darnell, R.B. *Nature* **460**, 479–486 (2009).
7. Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. *PLoS Biol.* **3**, e85 (2005).
8. Kishore, S. *et al. Nat. Methods* **8**, 559–564 (2011).
9. Chi, S.W., Hannon, G.J. & Darnell, R.B. *Nat. Struct. Mol. Biol.* **19**, 321–327 (2012).
10. Linsley, P.S. *et al. Mol. Cell. Biol.* **27**, 2240–2252 (2007).
11. Grimson, A. *et al. Mol. Cell* **27**, 91–105 (2007).
12. Leivonen, S.K. *et al. Oncogene* **28**, 3926–3936 (2009).
13. Selbach, M. *et al. Nature* **455**, 58–63 (2008).
14. Gennarino, V.A. *et al. Genome Res.* **19**, 481–490 (2009).
15. Grün, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C. & Rajewsky, N. *PLoS Comput. Biol.* **1**, e13 (2005).
16. Garcia, D.M. *et al. Nat. Struct. Mol. Biol.* **18**, 1139–1146 (2011).
17. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. *Nat. Genet.* **39**, 1278–1284 (2007).
18. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. *Genome Biol.* **11**, R90 (2010).
19. Miranda, K.C. *et al. Cell* **126**, 1203–1217 (2006).
20. Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. *RNA* **10**, 1507–1517 (2004).
21. Lorenz, R. *et al. Algorithms Mol. Biol.* **6**, 26 (2011).
22. Yang, J.H. *et al. Nucleic Acids Res.* **39**, D202–D209 (2011).

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

Inference of the MIRZA model. We defined a parameterized biophysical model to assign binding free energies to all possible miRNA-mRNA hybrid structures and quantify the binding affinity of different mRNA fragments to the RNA-induced silencing complex (RISC). Because a CLIP experiment does not provide accurate binding frequencies for all possible mRNA segments in the transcriptome but rather gives a set of fragments that are enriched relative to the expression of their mRNAs, we extract a set of highly enriched target sites, $m_1, m_2, \ldots m_n$ of standardized length M from the CLIP data, as described in "Argonaute 2 CLIP experimental data sets" below. We will make the assumption that the probability of obtaining a particular mRNA fragment m is proportional to the product of the abundance of the mRNA fragment and the fraction of time that the fragment is bound to a RISC. The latter quantity will depend on the binding free energy between the mRNA and RISC. Let $P(m|B)$ denote the 'background' abundance of mRNA fragment m in the transcriptome. Let $P(m|\mathrm{IP})$ denote the probability that when a single bound RISC is immunoprecipitated, this complex will contain a certain mRNA fragment m. This probability depends not only on the relative abundance of m but also on the relative abundances of the different miRNAs that can interact with the mRNA fragment in a RISC. Formally, the probability $P(m|\mathrm{IP})$ can be written as a sum over the probabilities $P(m,\mu|\mathrm{IP})$ that the immunoprecipitated fragment is bound to a RISC containing mature miRNA μ. If we denote by π_μ the fraction of all RISCs that are bound to some target site and that are guided by miRNA μ, then we have

$$P(m\,|\,\mathrm{IP}) = \sum_{\mu} P(m,\mu\,|\,\mathrm{IP}) = \sum_{\mu} P(m\,|\,\mu)\pi_\mu \qquad (1)$$

where $P(m|\mu)$ is the probability that a bound RISC containing miRNA μ is bound to fragment m.

The guide miRNA can form different hybrid structures with an mRNA fragment. Denoting individual hybrid structures by σ and the binding free energy of a RISC-embedded miRNA μ with mRNA fragment m in configuration σ by $E(\sigma,\mu,m)$, from the standard Boltzmann distribution of statistical physics we have that the fraction $P(m|\mu)$ of all RISCs that are loaded with miRNA μ and are bound in configuration σ to mRNA segment m is proportional to $e^{E(\sigma,\mu,m)}P(m|B)$ (note that we set the inverse temperature parameter β of the Boltzmann distribution to 1, for notational simplicity, which can be thought of as setting the scale of the energy parameters in units of $k \times T$, k being the Boltzmann constant and T the temperature). Thus, a RISC loaded with miRNA μ is bound to mRNA fragment m with probability

$$P(m\,|\,\mu) = \frac{\sum_{\sigma} e^{E(\sigma,\mu,m)} P(m\,|\,B)}{\sum_{m',\sigma'} e^{E(\sigma',\mu,m')} P(m'\,|\,B)} \qquad (2)$$

where the sum in the numerator is over all possible hybrid structures σ, and the sum in the denominator is over all possible hybrid structures and all possible M-nucleotide-long mRNA fragments m'. The probability of the entire data is

$$P(D) = \prod_{i=1}^{n} P(m_i\,|\,\mathrm{IP}) \qquad (3)$$

where the product is over all n mRNA fragments m_i that are sampled. The probability of observing a fragment m_i when randomly selecting fragments from the mRNA pool is just $P(m_i|B)$. Thus, the ratio of probabilities for observing the data under our model as opposed to random sampling is given by

$$R(D) = \prod_{i=1}^{n} \frac{P(m_i\,|\,\mathrm{IP})}{P(m_i\,|\,B)} = \prod_{i=1}^{n} R(m_i) \qquad (4)$$

The ratios $R(m_i)$ quantify to what extent the observation of m_i is explained by miRNA binding, i.e., they give the enrichment of fragment m_i when immunoprecipitating with a RISC relative to its abundance in the mRNA pool. Using equation (1) above we can write the enrichment of a fragment in terms of its enrichment for individual miRNAs

$$R(m) = \frac{P(m\,|\,\mathrm{IP})}{P(m\,|\,B)} = \sum_{\mu} \frac{P(m\,|\,\mu)}{P(m\,|\,B)}\pi_\mu = \sum_{\mu} R(m\,|\,\mu)\pi_\mu \qquad (5)$$

Target quality and target frequency. The quantity $R(m|\mu)$ represents the ratio of the probability that a RISC guided by miRNA μ binds to segment m and the background probability $P(m|B)$ of isolating segment m. In other words, $R(m|\mu)$ is the enrichment of fragment m among all fragments bound to a RISC loaded with miRNA μ relative to its background frequency $P(m|B)$. Because $R(m|\mu)$ quantifies the quality of segment m for miRNA μ (i.e., relative to all other possible target segments) we will refer to it as the target quality. Note, however, that for a given segment m, the miRNA with the highest target quality $R(m|\mu)$ is not necessarily the miRNA that most frequently associates with segment m because this latter quantity depends also on the relative abundances π_μ of RISCs that are loaded with different miRNAs. As can be seen from equation (5), the fraction of time that segment m is bound by miRNA μ, and, consequently, the miRNA that most frequently binds to segment m, is the one that maximizes the product $R(m|\mu)\pi_\mu$. We will refer to $R(m|\mu)\pi_\mu$ as the target frequency of miRNA μ for segment m.

Parameterization of the binding energies. Ignoring the possibility that the miRNA or the mRNA fragment form internal structures (base-pairing within themselves), our model assumes that each possible hybrid structure σ consists of one or more hybridized pairs of nucleotides that are separated by unpaired nucleotides, forming either symmetrical or asymmetrical loops, depending on whether the number of unpaired nucleotides in the miRNA and mRNA are the same or different. A hybrid σ can then be uniquely represented using the following set of 'moves': an initial hybridized pair (i,j), i.e., position i in the miRNA hybridized to position j in the mRNA fragment; addition of another hybridized pair immediately following the current pair; opening of a loop; addition of a symmetric pair of unhybridized nucleotides to the loop; addition of an unpaired nucleotide in the mRNA fragment; and addition of an unpaired nucleotide in the miRNA.

To ensure that each possible hybrid can be realized in only one way with these moves, we make the convention that asymmetric additions to loops can be followed only by more asymmetric additions of the same type or by a hybridized pair. Similarly, symmetric additions can be followed only by additional

© 2013 Nature America, Inc. All rights reserved.

symmetric additions, by an asymmetric addition or by a hybridized pair. Hybrids have to end in a hybridized pair, and the remaining nucleotides in the mRNA fragment and miRNA are considered dangling ends.

For each possible hybrid that can be constructed as described above, we assume that the binding energy can be decomposed into a structural and a sequence component:

$$E(\sigma, m, \mu) = E_{\text{struc}}(\sigma) + E_h(\sigma, m, \mu) \quad (6)$$

The structural contributions to the energy are determined from the moves and are an energy E_o for every loop that is opened, an energy E_{sym} for symmetrically extending a loop by one base in the miRNA and one base in the mRNA, an energy E_μ for asymmetrically extending a loop by an unpaired base in the miRNA, an energy E_m for asymmetrically extending a loop by an unpaired base in the mRNA fragment and an energy E_i when position i in the miRNA is hybridized. The latter reflects the constraints that the Argonaute protein imposes on the embedded miRNA, for example, through the accessibility of the corresponding position of the miRNA when it is inside a RISC. Without loss of generality, dangling bases in mRNA and miRNA per definition are assigned an energy $E_d = 0$. Thus, the structural part $E_{\text{struc}}(\sigma)$ depends on the number of loops, their sizes, their (a)symmetry and the positions in the miRNA that are hybridized. This dependency on miRNA positions enters through the energies E_i of the hybridized positions.

The sequence-dependent part of the energy consists of a sum of energy contributions for each hybridized pair, with $E_{\alpha\beta}$ being the energy contribution for hybridizing nucleotide α in the mRNA to nucleotide β in the miRNA. If we denote by h the set of miRNA positions that are hybridized in structure σ, we have

$$E_h(\sigma, m, \mu) = \sum_{i \in h} E_{m_i \mu_i} \quad (7)$$

with m_i being the nucleotide occurring at the position in the mRNA segment hybridized to miRNA position i, and μ_i the nucleotide at position i of the miRNA. Although in the most general case we would need 16 parameters to describe these contributions, we have considered only the usual base-pairing interactions A-U/U-A, C-G/G-C and G-U/U-G, which we described by parameters $E_{AU} = E_{UA}$, $E_{CG} = E_{GC}$, and $E_{GU} = E_{UG}$. We assign all other combinations a very negative energy, i.e. $-\infty$, such that they have zero probability of occurrence.

Removing redundancies of the parameterization. To infer the energy parameters from the observed data D, it is important to determine whether our parameterization contains redundancies, i.e., if there are global transformations of the parameters that would leave the overall likelihood ratio $R(D)$ invariant. In the model described above, a redundancy results from the fact that for every hybridized base pair (α,β), there is a sequence-dependent contribution $E_{\alpha\beta}$ and a structural contribution E_i from the hybridized position i in the miRNA. Thus, if we replace $E_{\alpha\beta}$ with $E_{\alpha\beta} + c$ (with c a constant) for all pairs (α,β) and at the same time replace E_i with $E_i - c$, then all energies $E(\sigma,m,\mu)$ remain unchanged. To remove this redundancy, we assign one of these parameters a neutral value. We chose to set $E_{GU} = 0$. The energies E_d of the dangling ends are set to 0 as well to avoid redundancies in the parameterization.

As detailed below, we fit all the energy parameters of the model by optimizing the likelihood of the observed CLIP data. The reader may wonder why certain parameters, such as the energies associated with base-pairing, are not simply set to experimentally estimated values such as those that are used in RNA secondary-structure prediction algorithms. It is important to stress that the energy parameters that we are inferring here are the effective contributions of various structural components (for example, base pairs and loops) in the context of the RISC. That is, the interaction of the miRNA and mRNA target will likely be strongly influenced by the context provided by this protein complex, and it is therefore not clear a priori what the contributions of different base pairs and loops should be.

Calculating target qualities and best hybrids. To infer the energy parameters, we search for the set of parameters that maximize the ratio $R(D)$ as given in equation (5); this requires calculating the target qualities $R(m|\mu)$, which in turn requires summing over combinatorially many hybrid structures σ, as in equation (2), i.e., performing partition sums. As detailed in the **Supplementary Note**, we have derived recursion relations that allow us to efficiently calculate these partition sums with standard dynamic programming techniques. In addition, as also detailed in the **Supplementary Note**, similar recursion relations can be used to determine the best hybrid structure for each pair of an mRNA fragment m and miRNA μ, i.e., the structure with the highest binding energy between fragment and miRNA.

Fitting the fraction of RISCs carrying specific miRNAs. Apart from all the energy parameters, the final likelihood ratio $R(D)$ also depends on the fractions π_μ of bound RISCs that are loaded with miRNA μ. As detailed in the **Supplementary Note**, given a set of energy parameters, the fractions π_μ that maximize the likelihood ratio $R(D)$ can be easily calculated using an expectation-maximization procedure. Thus, the numerical parameter-optimization procedure involves only the energy parameters.

Implementation of the parameter optimization. We implemented our MIRZA algorithm in C++, in an object-oriented framework. It takes as input FASTA-formatted files of mRNA fragments and miRNA sequences. To avoid biases introduced by the slight differences in length of different miRNAs, we trimmed all miRNA sequences to 21 nucleotides. We optimized the parameters of our biophysical model through simulated annealing, for which we used the GNU scientific library (http://www.gnu.org/s/gsl/) and an object-oriented library for numerical programming in C++ (O_2scl, http://o2scl.sourceforge.net/). For efficiency, we further used the Open Multi-Processing architecture (OpenMP, http://openmp.org/wp/), which supports multiplatform shared-memory parallel programming in C/C++ and Fortran. The parameters that we optimized were the base-pairing energies E_{AU} and E_{CG}, the loop energies E_o, E_{sym}, E_μ, E_m and the positional hybridization energies E_i, where $i = 1, 2, \ldots 21$.

For both the synthetic and Ago2 CLIP data sets, we performed multiple simulated annealing runs starting from random initial conditions, and we analyzed the reproducibility of the fitted parameters (see main text and **Supplementary Note**).

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.2341

Argonaute 2 CLIP experimental data sets. Of the recently reported data sets of Argonaute 2 binding sites, those generated with PAR-CLIP (photoactivatable ribonucleoside–enhanced cross-linking and immunoprecipitation) exhibit frequent diagnostic mutations (transition of uridine to cytidine), typically in the center of the miRNA-target site hybrid[8]. Within each single-linkage cluster that contained sites from at least three of four Ago2-CLIP or PAR-CLIP samples from ref. 8 that were generated with various protocols (GEO database accessions GSM714642, GSM714644, GSM714645 and GSM714647), we identified the nucleotide with the highest frequency of cross-link diagnostic mutations and extracted regions of 51 nucleotides centered on the position of cross-link (cross-link–centered regions, CCRs). A set of 2,988 high-confidence CCRs that were among the top 3,000 both in terms of coverage by sequence reads and in terms of enrichment in the read coverage relative to the read coverage of the same region in HEK293 mRNA-seq samples were retained for further analyses (**Supplementary Table 1**).

miRNA transfection data for functional analysis of predicted sites. To investigate the functionality of canonical and noncanonical targets predicted by various methods, we used published microarray data sets of changes in gene expression following the transfection of different miRNAs. We selected data sets corresponding mostly to miRNAs that are expressed in HEK293 cells from which CLIP data have been obtained. We further retained data from successful transfection experiments, meaning those in which the mRNAs carrying canonical sites for the transfected miRNA in their 3′ UTRs were significantly downregulated as compared to the other remaining mRNAs (Wilcoxon's rank-sum test on $\log_2$ fold changes, *P* value cutoff of 0.001) and discarded the other data sets. The five data sets that we thus used are summarized below.

In a first study[10], 11 miRNAs (miR-16, miR-15a, miR-106b, miR-20a, miR-103, miR-17, miR-20a and let-7c) were transfected in HCT116 and DLD-1 cell lines, each in duplicate. The processed differential expression data from the GEO database (accession GSE6838, experiments GSM156532, GSM156541, GSM156543, GSM156544, GSM156545, GSM156546, GSM156549, GSM156550, GSM156553, GSM156554, GSM156555, GSM156556, GSM156557, GSM156558, GSM156576 and GSM156580) together with the probe to transcript mapping provided by the authors as a SOFT-formatted file were downloaded. Probes associated with RefSeq transcripts according to the annotation were kept for subsequent analysis. Differential expression at the gene level was obtained by mapping RefSeq IDs to Entrez Gene IDs using the RefSeq database downloaded on 11 January 2007. For each gene, fold changes were averaged over the duplicate experiments.

In the second study[11], nine miRNAs (miR-122, miR-128, miR-132, miR-133a, miR-142-3p, miR-148b, miR-181a, miR-7 and miR-9) were transfected in HeLa cells, and mRNA expression was profiled 12 h and 24 h post-transfection. We retrieved the processed differential expression data from GEO (GSE8501) and then applied the same analysis performed on the previous data set to the fold-change data from the 24-h time point.

In the third study[12], miR-18a, miR-193b, miR-302c and miR-206 were transfected into MCF7 cells. We again retrieved the processed data from the GEO database (GSE14847) and computed average expression levels per Entrez Gene ID. We then computed the $\log_2$ fold change in expression levels upon miRNA transfection as compared to scrambled pre-miR control.

Another set of miRNA transfections in HeLa cells[13] involved miR-155, miR-16, miR-1, miR-30a and let-7b. We downloaded the CEL files from http://psilac.mdc-berlin.de/download/. Of these five miRNA transfections in HeLa, we excluded the let-7b experiment because of the reported negative feedback of let-7b on the RNAi pathway due to direct targeting of Dicer[13].

Finally, we obtained the CEL files of the miR-26b and miR-98 overexpression in HeLa cells[14] from GEO (accession GSE12100).

We imported the CEL files into the R software (http://www.R-project.org/) using the Bioconductor "affy" package[23]. The probe intensities were corrected for optical noise, adjusted for nonspecific binding and quantile-normalized with the gcRMA algorithm[24]. 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 $\log_2$ fold changes to obtain an expression change per gene.

All together, these five data sets cover changes in gene expression in 38 different transfection experiments involving 26 distinct miRNAs.

Comparison of miRNA target prediction methods. Some methods predict miRNA target sites, whereas others predict transcripts that are targeted by individual miRNAs. To be able to compare these heterogeneous predictions, we worked at the level of transcripts; for methods that predicted target sites, we assumed that the transcript score is given by the highest score of any predicted site in that transcript. The methods that we considered were ElMMo[3] (http://www.mirz.unibas.ch/ElMMo3/), which estimates the selection pressure on individual sites through comparative genomics; PicTar[15] (http://dorina.mdc-berlin.de/rbp_browser/hg18.html), another comparative genomics-based method whose predictions are widely used; TargetScan P_{CT}[2] (http://www.targetscan.org/), another evolutionary conservation–based method that, on the basis of previous evaluations, is considered one of the most accurate methods for identifying functional target sites; TargetScan context+[16], which predicts miRNA target sites according to the sequence context in which they occur in their host transcripts; miRanda[18] (http://www.microrna.org/microrna/getDownloads.do), a method that in its current version, mirSVR, uses support-vector regression based on a list of features of both the miRNA and its putative target (miRanda provides four separate files of targets, depending on whether targets are filtered on the basis of mirSVR score and/or conservation; we used the "S" sets of targets that are filtered by score, irrespective of conservation); PITA[17], release of 31 August 2008 (we extracted predictions for the miRNAs of interest from the web site, http://genie.weizmann.ac.il/pubs/mir07/mir07_dyn_data.html, queried with default parameters), which computes an energy of interaction between miRNA and target site, taking into account the structural accessibility of the target site; RNAduplex[20], which computes the minimum free energy of hybridization between the two RNA strands (we downloaded RNAduplex as part of the ViennaRNA package from http://www.tbi.univie.ac.at/RNA/RNAduplex.html and applied it to the entire set of representative 3′ UTRs of human genes); RNAhybrid[21], which uses an approach similar to

© 2013 Nature America, Inc. All rights reserved.

RNAduplex (we downloaded RNAhybrid from the server hosted by the University of Bielefeld (http://bibiserv.techfak.uni-bielefeld.de/) and applied it to the entire set of representative 3′ UTRs of human genes); and rna22 (ref. 19), a method based on statistical overrepresentation of miRNA-complementary motifs (current genome-wide predictions of this method were obtained from the authors).

We further included lists of targets of miRNAs from the starBase database[22] (http://starbase.sysu.edu.cn/), which intersects Ago2-CLIP sites with miRNA target predictions by TargetScan, PicTar, miRanda, PITA and rna22. StarBase does not provide a default sorting of predicted sites but allows users to manipulate stringency parameters. We downloaded target lists with the default settings of the database and also with the most inclusive settings that maximize the total number of predicted sites.

Because different methods may use different transcript collections as a basis for their predictions, we decided to compare predictions at the level of Entrez genes. For ElMMo, PicTar, TargetScan, miRanda, PITA and rna22, we collected for each Entrez gene all transcripts associated with the gene and defined the target score as the highest score among all transcripts in the set.

For RNAhybrid and RNAduplex, we predicted miRNA-complementary sites transcriptome wide. For this purpose, we selected a representative 3′ UTR for each Entrez gene that had a RefSeq transcript in the 18 January 2011 release of the RefSeq database. We chose as the representative 3′ UTRs those that had the longest transcript among those that were represented in RefSeq, had an annotated 5′ UTR, 3′ UTR and CDS and were associated with the corresponding gene. We scanned each 3′ UTR with windows of 50 nucleotides, shifting by 25 nucleotides at a time, and predicted the minimum free energy of interaction between the miRNA and each window. We then defined the transcript score as the minimum free energy over all windows from the 3′ UTR of a given transcript. For MIRZA, the target score of a transcript from the representative set was defined as the sum of the logarithms of the target qualities of all sites occurring in the transcript.

Median fold changes. To test the accuracy of the target predictions of each method, we used the data sets of miRNA transfection experiments as described in the section "miRNA transfection data for functional analysis of predicted sites." For each transfection experiment, and each method, we sorted all predicted target genes by score and filtered out all genes for which no fold change data were available in the corresponding data set. We then determined, as a function of the number n of top predicted targets, the median log fold change $\mathrm{lm}(n)$ of these targets in response to miRNA transfection. Lower median fold changes thus indicate that a method predicts targets that are more strongly downregulated upon transfection of the miRNA. For each of the five data sets, we calculated average median log fold changes $\langle \mathrm{lm}(n) \rangle$ by averaging the functions $\mathrm{lm}(n)$ over the transfection experiments in individual data sets. We also calculated an average over all 38 transfection experiments.

Estimating the number of functional targets. Besides calculating median log fold changes, we also determined—for each miRNA and each method—the fraction $f(n)$ of the top n predicted targets that were downregulated as a function of the number n of top predicted targets. We used the functions $f(n)$ to estimate the total number of functional targets as follows. For each data set, we first determined the total fraction f_{tot} of downregulated transcripts among all transcripts for which fold change data were measured. Typically, f_{tot} is close to 50%. Thus, if we were to make random predictions, we expect a fraction f_{tot} of the predicted targets to be downregulated. If $f(n)$ is considerably larger than f_{tot}, this indicates that there must be true targets among the n predicted targets. Note that, if a fraction $\rho(n)$ of the n predicted targets are true targets, and using the fact that true targets must be downregulated per definition, then the total fraction $f(n)$ of downregulated targets will be $f(n) = \rho(n) + f_{\mathrm{tot}}\,(1 - \rho(n))$. From this we can estimate the total number of functional targets as $n_{\mathrm{func}}(n) = n \times \rho(n) = n\,(f(n) - f_{\mathrm{tot}})/(1 - f_{\mathrm{tot}})$. For each method and each transfection experiment, we determined the total number of functional targets by maximizing $n_{\mathrm{func}}(n)$ over n: i.e., we chose the number n of top predicted targets such that $n_{\mathrm{func}}(n)$ is maximal $n_{\mathrm{func}} = \max_n[n_{\mathrm{func}}(n)]$. For each method, we also determined the average number of functional targets $\langle n_{\mathrm{func}} \rangle$ for each of the five data sets by averaging n_{func} over the transfection experiments in a data set. We also calculated an overall $\langle n_{\mathrm{func}} \rangle$ averaging over all 38 transfection experiments. Finally, all these calculations were also performed in a way that restricted the targets to those transcripts that do not contain a canonical match to the seed sequence of the miRNA, as described in the next section.

Noncanonical binding sites. To identify noncanonical target sites among the CLIP sites, we used the following stringent procedure. We first predicted with MIRZA the miRNA μ with which each mRNA fragment m most likely interacted, i.e., the miRNA for which the mRNA fragment had the highest target frequency $R(m|\mu)\pi_{\mu}$. Next we determined the optimal hybrid σ for this miRNA-mRNA target pair with the recursion relations described in the **Supplementary Note**, and on the basis of these hybrids we divided the set of mRNA fragments into two subsets: canonical sites, which base-paired contiguously with nucleotides 2–8 of the miRNA or had an exact match to positions 2–7 of the miRNA followed by an adenine (which would be positioned opposite position 1 of the miRNA), and noncanonical sites, for which the above condition was not satisfied.

We then identified transcripts that contained a single, noncanonical CLIP–cross-linked site for the transfected miRNA and retained those transcripts that did not additionally contain a canonical seed match (as defined above) anywhere in the 3′ UTR. We used in this search the 3′ UTRs of representative transcripts from ref. 8. This procedure gave us a conservative set of transcripts on which the miRNA was likely to act on a noncanonical site. We sorted the noncanonical sites according to their target quality $R(m|\mu)$ with respect to the transfected miRNA μ, and we then divided the set into three subsets of equal size, corresponding to the top 33%, the middle 33% and the bottom 33% in terms of the target quality. To resolve issues of differences between genome and transcriptome annotations, we investigated the change in expression at the level of genes. That is, we mapped transcripts to corresponding genes in the Entrez database of NCBI. Finally, we compared the expression-level changes between genes containing sites within each subset and genes whose representative transcripts did not contain a seed match in the 3′ UTR or did contain a seed match (irrespective of whether it was CLIP

© 2013 Nature America, Inc. All rights reserved.
npg

doi:10.1038/nmeth.2341

cross-linked) in the 3′ UTR. For each gene, we computed the average log fold change across replicate transfection experiments.

For the comparison of prediction accuracy of the different target prediction methods, we defined noncanonical targets as follows. For each miRNA, we scanned all 3′ UTRs of RefSeq transcripts associated with each Entrez gene for a canonical match to the miRNA. All Entrez genes for which such a seed match was detected are considered canonical targets by default, i.e., irrespective of where in the 3′ UTR the various methods predicted sites or which of the RefSeq transcripts contained such a site. Thus, noncanonical target genes of a given miRNA are those for which the 3′ UTRs of associated RefSeq transcripts do not contain a canonical match to the seed sequence.

Representation of noncanonical binding modes among CLIP sites. To determine the prevalence of specific noncanonical binding 'modes' in CLIP data sets, we extracted sites as follows. From each of the four Ago2 data sets from ref. 8, and from the three mouse brain Ago2 HITS-CLIP data sets (libraries prepared from the 130-kDa band) in ref. 6, we extracted the 5,000 sites with the highest coverage by reads. We also extracted the 5,000 most enriched sites, relative to the expression of the corresponding mRNAs in an mRNA-seq sample that we prepared from HeLa cells, from the two samples from ref. 6 that were obtained after miR-124 transfection in HeLa cells. We applied the MIRZA model to each of these sets of putative Ago2 binding sites to determine the miRNAs that most likely guided the interaction with the site and the hybrid with the highest score, and we used this to determine the relative proportions of individual binding modes (for example, that with a bulge at the pivot position[9]) among these hybrids.

23. Gentleman, R.C. *Genome Biol.* **5**, R80 (2004).
24. Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. *J. Am. Stat. Assoc.* **99**, 909–917 (2004).

© 2013 Nature America, Inc. All rights reserved.